Vol. 31, No. 4, 2018

CHANCE

Using Data to Advance Science, Education, and Society



Epic Fail? The Polls and the 2016 Presidential Election

Including...

Elementary Statistics on Trial—The Case of Lucia de Berk

09332480(2018)31(4)

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 Graphics

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.

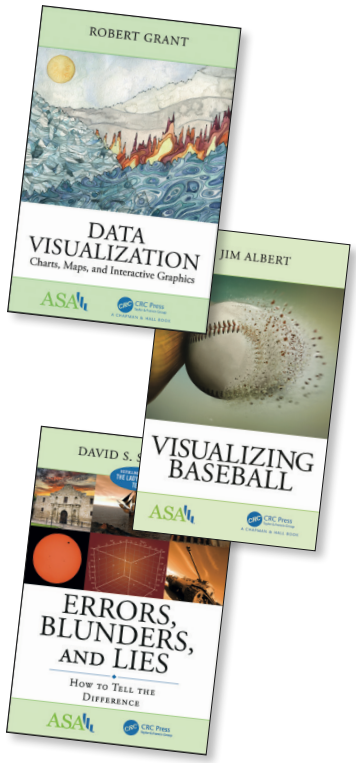
August 2017 • 142 pp • Pb: 9781498782753: ~~$29.95~~ $23.96 • www.crcpress.com/9781498782753

ERRORS, BLUNDERS, AND LIES
How to Tell the Difference

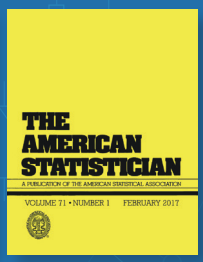
David S. Salsburg, Emeritus, Yale University, New Haven, CT, USA

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

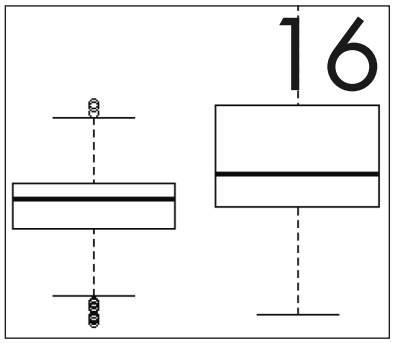
http://bit.ly/CRCASA2018

CHANCE

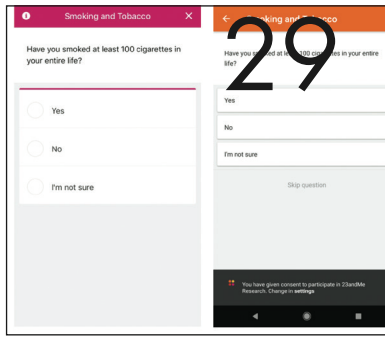
Using Data to Advance Science, Education, and Society

http://chance.amstat.org









ARTICLES

4 Epic Fail? The Polls and the 2016 Presidential Election
Arnold Barnett

9 Elementary Statistics on Trial—The Case of Lucia de Berk
Richard D. Gill, Piet Groeneboom, and Peter de Jong

16 A Simple Example of Isolation in Building a Natural Experiment
Jose R. Zubizarreta, Dylan S. Small, and Paul R. Rosenbaum

24 "With Great Power Comes Great Responsibility": Harnessing the Extraordinary Power of Statistics
Daniel H. Mowrey, Jonathan Potts, Susan Spruill, Walter Stroup, and Michiko I. Wolcott

29 Mobile Apps versus Web Browsers: A Comparison of Self-administered Survey Platforms
Briana Cameron, 23andMe Research Team, and Robert Gentleman

37 The Lunacy of Motorcycle Mortality
Donald A. Redelmeier and Eldar Shafir

45 Meta-Analysis: Can We Extract Gold from the Biomedical Literature?
Deborah V. Dawson and Derek R. Blanchette

COLUMNS

51 Taking a Chance in the Classroom
Dalene Stangl and Mine Çetinkaya-Rundel, Column Editors

Top 10 Reasons Not to Drop Insignificant Regressors: A Statistical Listicle
Chris Rohlfs

DEPARTMENTS

3 Editor's Letter

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: Getty Images

EXECUTIVE EDITOR

Scott Evans
Harvard School of Public Health, Boston, Massachusetts
evans@sdac.harvard.edu

ADVISORY EDITORS

Sam Behseta
California State University, Fullerton

Michael Larsen
St. Michael's College, Colchester, Vermont

Michael Lavine
University of Massachusetts, Amherst

Dalene Stangl
Carnegie Mellon University, Pittsburgh, Pennsylvania

Hal S. Stern
University of California, Irvine

EDITORS

Jim Albert
Bowling Green State University, Ohio

Phil Everson
Swarthmore College, Pennsylvania

Dean Follman
NIAID and Biostatistics Research Branch, Maryland

Toshimitsu Hamasaki
Office of Biostatistics and Data Management
National Cerebral and Cardiovascular Research Center, Osaka, Japan

Jo Hardin
Pomona College, Claremont, California

Tom Lane
MathWorks, Natick, Massachusetts

Michael P. McDermott
University of Rochester Medical Center, New York

Mary Meyer
Colorado State University at Fort Collins

Kary Myers
Los Alamos National Laboratory, New Mexico

Babak Shahbaba
University of California, Irvine

Lu Tian
Stanford University, California

COLUMN EDITORS

Di Cook
Iowa State University, Ames
Visiphilia

Chris Franklin
University of Georgia, Athens
K-12 Education

Andrew Gelman
Columbia University, New York, New York
Ethics and Statistics

Mary Gray
American University, Washington, D.C.
The Odds of Justice

Shane Jensen
Wharton School at the University of Pennsylvania, Philadelphia
A Statistician Reads the Sports Pages

Nicole Lazar
University of Georgia, Athens
The Big Picture

Bob Oster, University of Alabama, Birmingham, and **Ed Gracely**, Drexel University, Philadelphia, Pennsylvania
Teaching Statistics in the Health Sciences

Christian Robert
Université Paris-Dauphine, France
Book Reviews

Aleksandra Slavkovic
Penn State University, University Park
O Privacy, Where Art Thou?

Dalene Stangl, Carnegie Mellon University, Pittsburgh, Pennsylvania, and **Mine Çetinkaya-Rundel**, Duke University, Durham, North Carolina
Taking a Chance in the Classroom

Howard Wainer
National Board of Medical Examiners, Philadelphia, Pennsylvania
Visual Revelations

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.

Copyright ©2018 American Statistical Association. All rights reserved. No part of this publication may be reproduced, stored, transmitted, or disseminated in any form or by any means without prior written permission from the American Statistical Association. The American Statistical Association grants authorization for individuals to photocopy copyrighted material for private research use on the sole basis that requests for such use are referred directly to the requester's local Reproduction Rights Organization (RRO), such as the Copyright Clearance Center (*www.copyright.com*) in the United States or The Copyright Licensing Agency (*www.cla.co.uk*) in the United Kingdom. This authorization does not extend to any other kind of copying by any means, in any form, and for any purpose other than private research use. The publisher assumes no responsibility for any statements of fact or opinion expressed in the published papers. The appearance of advertising in this journal does not constitute an endorsement or approval by the publisher, the editor, or the editorial board of the quality or value of the product advertised or of the claims made for it by its manufacturer.

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,*

Scott Evans

Epic Fail? The Polls and the 2016 Presidential Election

Arnold Barnett



Donald Trump's victory in the 2016 presidential election caused widespread shock, in large part because political polls seemed to predict an easy victory for Hillary Clinton. *New York* magazine was so confident of a Clinton victory that, the week of the election, its cover featured the word "loser" stamped across Donald Trump's face.

The day after the election, *Politico* asked, "How Did Everyone Get It So Wrong?" The next month, a columnist for the *New Republic* asked, "after 2016, can we ever trust the polls again?" given that they "failed catastrophically" in the presidential race. A *Wall Street Journal* columnist doubtless referred to the 2016 outcome when he noted that recent election polls had been "spectacularly wrong."

Was the primary failure among the polls or, instead, among Americans who, for whatever reason, were unable to recognize what the polls were actually saying? Here we explore the issue by considering two primary consolidators of presidential polls: FiveThirtyEight.com, the Nate Silver website that was previously considered the gold standard among consolidators, and Real Clear Politics, the immensely popular political website that had 32 million unique visitors in the month before the 2016 election.

It is of interest to know what these consolidators reported on the eve of the election, and then consider what their forecasts implied about polling accuracy in key states that were credited or blamed for Hillary Clinton's "reversal of fortune."

Properly understood, the publicly available polls suggested that the Clinton/Trump race was what the media, in comparable circumstances, would typically call a "statistical dead heat." The belief that the Democrat was all but certain to win was inconsistent with both a superficial awareness of what the polls said and a deeper understanding of what they implied about the outcome.

To the question "Can we ever trust the polls again?" after what happened in 2016, the answer is an emphatic yes.

On Election Day

Even a cursory look at the FiveThirtyEight or Real Clear Politics on the morning of Election Day 2016, would have suggested that a Clinton victory was far from inevitable. In its poll-based forecast, FiveThirtyEight estimated then that Trump had a 28.6% chance of winning the presidency. Real Clear Politics predicted that the Electoral College breakdown of the results would be 272 for Clinton and 266 for Trump.

That prognosis implied that, if only one state switched from Clinton to Trump, Trump would be president. Even if the switching state had only three electoral votes (the minimum number), the result would be a 269–269 tie, in which case, Trump would prevail when the winner was chosen by the Republican House of Representatives.

To be sure, on Election Day itself, many people were no longer looking at polls, but polling outcomes a few days before the election were, if anything, more favorable to Trump: On November 4, 2016, for example, FiveThirtyEight estimated Trump's chances of winning four days hence at 35.4%. Thus, evidence of a close race cannot be said to have arisen too late to be widely noticed.

Paradoxically, respect for the two consolidators and the state-specific polls they summarized actually grows if one looks further at what they supposedly did wrong.

FiveThirtyEight's "Errors" in Three Key States

The great surprise in the 2016 election was that Donald Trump carried Pennsylvania, Michigan, and Wisconsin, three states that were considered part of the Democratic "firewall." FiveThirtyEight did not contradict the general impression: Its polls-only forecast the morning of the election was that Trump had a 21.1% chance of winning Michigan, a 23.0 chance of winning Pennsylvania, and a 16.5% chance of carrying Wisconsin. (Those "polls-plus" forecasts that also considered economic and historical factors were essentially the same.)

A simple calculation with the FiveThirtyEight estimates might suggest that Trump's chances of achieving a "trifecta" in these states was essentially .211 * .23 * .165 = 1 in 125, but such a calculation would be unfair to FiveThirtyEight.

Advancing the statistic "1 in 125" is misleading for two reasons. The first is that judging a forecasting model by what emerged as its weakest forecasts is systematically biased against it. The second is that the calculation treats the three outcomes as independent, when FiveThirtyEight explicitly assumes that they are not. Although its specific procedures are not transparent, FiveThirtyEight recognized that the outcomes in these three similar states are correlated. Indeed, on election night, TV commentators readily grasped that Trump's strong performance in Pennsylvania increased the chance that he would also carry Michigan and Wisconsin.

A simple analysis indicates the strength of the pattern of correlation. Assuming that FiveThirtyEight's probability assessments were accurate in Pennsylvania and Wisconsin (at 23.0% and 16.5% for Trump, respectively), how might one estimate the chance that he would win in both states?

One reflection of the correlation is that, in the 12 elections before 2016, Pennsylvania and Wisconsin went for the same candidate 10 times (whether Republican or Democrat, and whether the national winner or the national loser). Using this statistic, we can set up some simple equations that approximate the chance that Trump would carry the two states.

Some definitions:

P_{CC} = probability that Clinton would win both Pennsylvania and Wisconsin

P_{TT} = probability that Trump would win both Pennsylvania and Wisconsin

P_{CT} = probability that Clinton would win Pennsylvania and Trump would win Wisconsin

P_{TC} = probability that Trump would win Pennsylvania and Clinton would win Wisconsin

On the morning of the election, we could write:

$P_{CC} + P_{TT} = 10/12$ (based on the historical chance both states go the same way)

$P_{TT} + P_{TC} = .230$ (based on FiveThirtyEight's forecast of the chance Trump would win Pennsylvania)

$P_{TT} + P_{CT} = .165$ (based on FiveThirtyEight's forecast of the chance Trump would win Wisconsin)

$P_{CC} + P_{TT} + P_{CT} + P_{TC} = 1$ (because something has to happen)

Solving this system of linear equations provides:

$P_{TT} = .114$ $P_{TC} = .116$ $P_{CT} = .051$ $P_{CC} = .719$

That $P_{TT} = .114$ (as opposed to .23 * .165 = .038) implies that the tendency of Pennsylvania and Michigan to "go the same way" yields a much greater chance of a double victory than would arise under the FiveThirtyEight projections and an independence assumption. The chance that Trump would carry at least one of the states would be estimated by 1 - .719 = .2.

What about Michigan, as well as Pennsylvania and Wisconsin? The historical pattern is that, over the 12 presidential elections from 1960 to 2012, Michigan went along with Pennsylvania and Wisconsin every time the latter two states voted for the same

Table 1—Three Estimates of the Probability that Trump Would Carry all of Michigan, Pennsylvania, and Wisconsin in 2016, and Carry at Least One

Initial States[a]	P (Trump Carry All Three)	P (Trump Carry at Least One)
PA, WI	.114	.281
PA, MI	.146	.396
MI, WI	.063	.313
Average	.108	.330

[a] Two states were used in the initial formulation with four linear equations and four unknowns, with the third state included via the frequency with which it supported the same presidential candidate as the other two when the other two agreed over 1960-2012.

candidate (i.e., 10 times out of 10). It therefore could be approximated:

$$P_{TTT} = \text{P(Trump would win PA, MI, and WI)} \approx P(\textit{win PA and MI}) * (\tfrac{10}{10}) = .114$$

and:

$$P(\textit{Trump win at least one of PA, MI, and WI}) \approx 1 - P(\textit{Clinton wins all three}).$$

$$\textit{or } P(\textit{Trump win at least one of PA, MI, and WI}) \approx 1 - .719 * (\tfrac{10}{10}) = .281$$

These calculations, however, do not actually use FiveThirtyEight's estimate that Trump's chance of carrying Michigan was .211. One could perform the analogous calculations based initially on Pennsylvania plus Michigan, and on Michigan plus Wisconsin, as shown in Table 1.

The key probabilities exhibit some variation over the three approximations, but the collective results suggest roughly a 1 in 9 chance that Trump would win all three states, and about a 1 in 3 chance that he would carry at least one (The use of an omnibus linear model that considers all three of FiveThirtyEight's projections at once—with eight equations and eight unknowns—does not yield more reliable probability estimates than the three-part approximation above.)

This 1 in 3 probability is especially noteworthy because, outside PA, MI, and WI, Trump amassed 260 electoral votes. That outcome was very much consistent with the state-by-state projections by both FiveThirtyEight and Real Clear Politics, but, once Trump had 260 electoral votes, he only had to win one of PA/MI/WI to achieve a national victory. If the chance of that outcome was something like 33%, it is no surprise that FiveThirtyEight saw a probability near 3 in 10 that Trump would triumph.

The Democratic "firewall" was far from fireproof.

These probability calculations are obviously approximate, and weigh historical patterns almost as much as the individual FiveThirtyEight projections, but the general point of the exercise survives its imperfections. There was reason to believe that, if Trump performed better-than-expected in one of the three states, he was more likely to do so in the others. Because FiveThirtyEight gave Trump an appreciable chance of winning each of the three states and recognized that the outcomes were not independent, it saw a Clinton victory as far from a sure thing—and it did so based largely on in-state polling results, which by no means deserved to be described as "spectacularly wrong."

Real Clear Politics

As noted, Real Clear Politics (RCP) predicted on Election Day that Trump would get 266 electoral votes, compared to Clinton's 272. It correctly identified the winner in 47 of the 51 states, erring—like FiveThirtyEight—in PA, MI, and WI, and also in Nevada. (FiveThirtyEight got Nevada right, but it tilted toward Clinton in Florida and North Carolina, which she lost.) However, RCP listed all four of the states it got wrong in its "tossup" category.

Still, the specific procedure by which RCP assigned a state's winner would cause statisticians to tear their hair out. Roughly speaking, it considered all major polls in the state within two weeks of the election, and took a simple arithmetical average of their results. If Trump was ahead of Clinton in this average, he was awarded the state in RCP's "no tossup" assessment.

The procedure obliterated the distinction between:

- A poll one day before the election and another 13 days before
- A poll with sample size 1,500 and another with sample size 400
- A poll among all registered voters and another restricted to "likely" voters

Table 2—Final 2016 RCP Poll Listings for the State of Ohio

Dates	Source	Clinton	Trump	Other	Sample Size
11/4–11/5	Emerson	39 (%)	46	15	900
11/2–11/4	CBS/YouGov	45	46	9	1188
11/1–11/2	Remington	44	45	11	2557
10/27–11/1	Quinnipac	41	46	13	589

Furthermore, RCP assigned no margin of error to its state-specific forecasts.

It is desirable to place the RCP results on a stronger statistical foundation, especially as an indicator of the caliber of state-level polls that it summarized. A simple way to do so is to accept RCP's assumption that the various polls in its average are essentially interchangeable. One can create a "megapoll" by combining all the individual polls RCP used for a given state in its Election Day forecast. Using that megapoll, one can estimate a margin of sampling error for the statewide result and—more importantly—estimate the probability that Trump would carry the state.

For example, the final 2016 RCP poll listings for the state of Ohio are shown in Table 2.

(It was not RCP that rounded-off polling results to the nearest percentage point; rather, that convention was followed by all four polls in their announcements of the results.)

Trump's average lead over Clinton in the four Ohio polls (weighed equally) was 14/4 = 3.5 percentage points, and that was the statistic featured by RCP in its summary about Ohio.

Although an average of 12% of those canvassed in the four polls were either undecided or supportive of a third-party candidate, the tacit assumption was that such voters could effectively be ignored in the race between Clinton and Trump.

The Emerson poll in Ohio generated 900 * .39 = 351 supporters of Clinton and 900 * .46 = 414 of Trump. For CBS/YouGov, the corresponding numbers were 546 and 535. For the four merged polls, Clinton had 2,252 supporters while Trump had 2,382. Obviously, the large Remington poll had a disproportionate effect on the merged result. If no relevance is accorded to the identity of the pollster, though, the number of days until the election, and the likelihood a polled voter will actually vote, then one can presumably construe the Clinton/Trump result in Ohio as arising from a random sample of 2,252 + 2,382 = 4,634 voters.

Among such voters, Trump's share of the Clinton/Trump vote was 51.4% and Clinton's was 48.6%.

Taking sampling error into account, what is the probability that Trump was actually ahead in Ohio, given the result of the "megapoll?" Starting with a uniform before Trump's vote share versus Clinton and revising it in Bayesian manner with the findings for the 4,634 voters sampled, the chance that Trump would defeat Clinton in Ohio would be estimated as .972. (This result arises from a beta distribution with α = 2,383 *and* β = 2,253.)

Proceeding similarly, one can get a poll-based estimate of the chance that Trump would win in each of the 51 states.

An RCP Projection

If V_i is state i's number of electoral votes (i = 1, .., 51) and the indicator variable X_i is defined by:

$$X_i = \begin{cases} 1 \text{ if Trump wins state } i \\ 0 \text{ if Trump loses state } i \end{cases}$$

then T—Trump's total number of electoral votes—follows:

$$T = \sum_{i=1}^{51} V_i X_i$$

Then, if $P_i = P(\textit{Trump will win state } i \mid \textit{recent polling results there})$, the mean of T is given by:

$$E(T) = \sum_{i=1}^{51} V_i E(X_i) = \sum_{i=1}^{51} V_i P_i$$

RCP never mentions possible correlations among the X_i's, and tacitly treats the X_i's as independent (implying that similarities across states are already appropriately reflected in the point estimates of the X_i's). Under independence, the variance of T would approximately follow:

$$\sigma^2(T) = \sum_{i=1}^{51} V_i \ \sigma^2(X_i) \approx \sum_{i=1}^{51} V_i P_i(1 - P_i)$$

To estimate $P(\textit{Trump wins}) = P(T \geq 269 \textit{ out of FiveThirtyEight Electoral Votes})$, we might be tempted to assume that T is normally distributed as the sum of 51 independent random variables. However, the heavy majority of P_i's (calculated with the method

shown above for Ohio) are very close to either zero or one. As a practical matter, therefore, the uncertainty in T depends on a small subset of the state-specific variables.

Thus, to approximate $P(Trump\ wins)$, we turn to simulation rather than treat T as normal, using the calculated P_i's = 258.8, with a standard deviation of about 17.3.

More than 2,000 simulation runs using the state-specific P_i's, the fraction in which Trump was victorious was .290. This estimate of $P(Trump\ wins)$ might seem smaller than expected given RCP's estimated 266–272 split. But, as the value of $E(T)$ suggests, Trump was only modestly ahead in several states assigned to him (i.e., several of his P_i's only moderately exceeded 1/2), while Clinton's leads were more substantial in the states where she led.

What should we make of the estimate that $P(Trump\ wins) \approx .290$ (which, coincidentally, is very similar to the estimate made by FiveThirtyEight)? Consider a pre-election poll that suggests that candidate A leads candidate B by 51% to 49%, with a margin of error of 3 percentage points. The media would routinely describe the race as "a statistical dead heat" or at least "too close to call." If, on election day, candidate B were to win with 50.3% of the vote, few people would cite the result as evidence that the poll was wrong.

This point is relevant because, under the standard formula for a 95% confidence interval, the sample size in this hypothetical poll with its 3-percentage-point error margin would be 1,068, and the split corresponding to 51%/49% would be 544 for A and 524 for B. Under a Bayesian revision of a uniform prior, B's true share of the vote would follow a beta distribution with $\alpha = 525$ *and* $\beta = 545$), and the chance that B was actually ahead would be estimated as 25.0%.

An obvious question arises: If a poll is not seen as inaccurate when it had assigned the actual winner a 25.0% chance of victory, how can one say that RCP was "spectacularly wrong" when it implied a 29.0% probability that Trump would win the presidency?

Indeed, in a two-candidate race with no undecided voters and a poll with a 3-percentage-point margin of error, the outcome would be treated as "too close to call" if B's estimated vote share was anything above 47%. But that means that the race would be viewed as undecided even when B's approximate chance of actually winning was as low as .03. The observed split corresponding to a 29% chance that B would win would be 50.8% for A and 49.2% for B, and the split—which obviously suggests a close race—is statistically equivalent to the result implied by Real Clear Politics in 2016.

The good performance of RCP—despite its ignoring the correlation among outcomes in PA, MI, and WI—is quiet testimony to the high accuracy of the individual polls it synthesized.

If the conventional wisdom is that the polls failed in 2016 systematically, then the conventional wisdom is somewhat more in error than the polls were.

In Conclusion

The day after the election, the *New York Times* ran the headline "Donald Trump's Victory Is Met with Shock Across a Wide Political Divide." That assessment was clearly accurate but, as we have suggested, this "shock" should be viewed as arising less because of the polls than despite them. Ironically, the real lesson from 2016 could be that not that we should pay less attention to the polls, but that we should pay more attention.

Further Reading

Caldwell, C. 2017. France's Choice: Le Divorce? *Wall Street Journal*.

FiveThirtyEight. 2016. Who will win the presidency? Available at *https://projects.fivethirtyeight.com/2016-election-forecast/*.

Healy, P. , and Peters, J. 2016. Donald Trump's Victory Is Met with Shock Across a Wide Political Divide. *New York Times*.

Narea, N. 2016. After 2016, Can We Ever Trust the Polls Again? *New Republic*.

Real Clear Politics, 2016 Presidential Election, No Toss-Up States. Available at *https://www.realclearpolitics.com/epolls/2016/president/2016_elections_electoral_college_m ap_no_toss_ups.html*.

Vogel, K., and Isenstad, A. 2016. How Did Everyone Get It So Wrong? *Politico*.

Author Bio

Arnold Barnett is the George Eastman professor of management science and professor of statistics at MIT's Sloan School of Management. He has received the President's Award for "outstanding contributions to the betterment of society" from the Institute for Operations Research and the Management Sciences (INFORMS), and the Blackett Lectureship from the Operational Research Society of the UK. He has also received the Expository Writing Award from INFORMS and is a fellow of that organization. He has been honored 15 times for outstanding teaching at MIT.

Elementary Statistics on Trial—The Case of Lucia de Berk

Richard D. Gill, Piet Groeneboom, and Peter de Jong

The trial of the Dutch nurse Lucia de Berk, suspected of several murders and attempted murders, was a high-profile case in the Netherlands. The initial suspicion rested mainly on quasi-statistical considerations, which produced (based partly on incorrect calculations) extremely small probabilities. Since the outcomes proved controversial, the court claimed to have dropped the statistical calculations from the verdict, but the verdict still rested on intuitive notions as "very improbable." That put statistics at center stage.

In the conviction of de Berk, a simple (so-called) hypergeometric model played an important role. The law psychologist Henk Elffers, consulted as a statistician by the court, used it, and it produced very small probabilities of occurrences of certain numbers of incidents. If we take into account the variation among nurses in incidents they experience during their shifts, these probabilities become considerably larger. This points to the danger of using an over-simplified discrete probability model in such circumstances.

The outcomes of applying our alternative model to this case are in striking contrast with those of the first calculations that led to the initial suspicions and were instrumental in determining the atmosphere surrounding the trial and subsequent hysteria. The main result is that under the assumption of heterogeneity, the probability of experiencing a number of incidents (14) that led to de Berk's conviction is about 0.020616 or one in 49 if the calculations are based on the same data as Elffers used. In his calculation, however, this probability was equal to 1 in 342 million.

The Data

The data for this discussion are from Elffers's unpublished reports. Before going into this, some general remarks about the data collection are needed.

One of the key features of the data was the flawed collection. Here, different disciplines came into conflict: Criminal investigation and scientific data gathering are very different. Their objectives, methods, and results are not compatible. Criminal investigation begins when there is suspicion of a crime; one is looking for or hunting down a suspect. If there is a need for meaningful statistics, the methodology must guarantee clear definitions and uniformity of the data collection.

In the case of Lucia de Berk, this clash of cultures proved disastrous. Incidents outside de Berk's shifts were discarded, and some initially reported incidents were later relabeled without clear reasons. Extra shifts without incidents and incidents outside shifts that de Berk worked were subsequently brought to light. Moreover, the data collection rested, for a large part, on memory.



Lucia de Berk

Clearly, the context of a criminal investigation produces a specific mindset: On the one hand, the witnesses know what is looked for (and some of them may already be convinced of the guilt of the suspect); on the other hand, a fear of implicating oneself and friends can considerably distort memory. The data on shifts and incidents for the period that were singled out in Elffers's reports are shown in Table 1 (this version of the table corrects an error in Meester, Collins, Gill, and van Lambalgen, where the total number of shifts in the ward RCH-41 was given as 336 instead of 366 in the table at the top of their page 235).

Elffers's Method

Analysis by Elffers, the statistician consulted by the court, was based on Table 1. As was noticed later, de Berk had actually worked three

Table 1—Data on Shifts and Incidents

Hospital name (and ward number)	JCH	RCH-41	RCH-42	Total
Total number of shifts	1,029	366	339	1,734
de Berk's number of shifts	142	1	58	201
Total number of incidents	8	5	14	27
Number of incidents during de Berk's shifts	8	1	5	14

JCH and RCH denote the "Juliana Children's Hospital" and "Red Cross Hospital," respectively, and 41 and 42 were different ward numbers of the Red Cross Hospital.

shifts in RCH-41 instead of just one, but this calculation is based on the data used by Elffers. Elffers argued by conditioning on part of the data and used two fundamental assumptions:

1. There is a fixed probability p for the occurrence of an incident during a shift (for example, p does not depend on whether the shift is a day shift or a night shift or on the nurse involved, etc.),
2. Incidents occur independently of one another.

On the basis of these assumptions, one can compute the probability that L incidents occur during de Berk's shifts, given the total number I of incidents and the total number N of shifts considered in the period of study. This is a hypergeometric probability, given by

$$\frac{\binom{S}{L}\binom{N-S}{I-L}}{\binom{N}{I}}$$

where S is the number of shifts de Berk worked and I is the total number of incidents, and where $\binom{s}{l}$, etc., denote binomial coefficients. Taking all the data of Table 1 together gives a total number of $N = 1{,}734$ shifts. Where de Berk had $S = 201$ shifts, there was a total number $I = 27$ of incidents, and $L = 14$ incidents during de Berk's shifts. If we evaluate this equation with these values for N, S, I, and L, we get the very small probability of about 1 in 4.2 million.

Computing the probability (p-value) that a nurse is present with 14 or more incidents in Elffers's method of testing a null hypothesis of no systematic effects on these combined data (but he actually did not test it in this way on the combined data; see below) requires summing the probabilities for $L = 14, 15, \ldots, 27$, for the probability of about 1 in 3.8 million. This is a very small probability, although still about 100 times larger than the probability Elffers arrived at as described below. For the model introduced in the next section, however, which used the same data, there is a probability of 1 in 49.

Elffers proceeded somewhat differently—not combining the data of the different hospitals (the details of what he actually did are described in Meester, Collins, Gill, and van Lambalgen). The most-important mistake he made in his calculation was to take the three hospitals separately, and multiply the probabilities he got for these separately. This has the absurd consequence that a nurse working in several different hospitals gets a higher chance of being accused of inexplicably being present at incidents than a nurse working in just one hospital.

In this way, he arrived at his estimate that the probability that Lucia de Berk was present at the given number of incidents at the Juliana Children's Hospital and the Red Cross Hospital was equal to 1 in 342 million. (Meester, Collins, Gill, and van Lambalgen, "Math error number 7: the incredible coincidence" [Schneps and Colmez]).

Post-hoc Testing

There is a danger of post-hoc testing: testing a hypothesis using the same data that suggested this hypothesis. Elffers actually tried to take this problem into account by starting from the assumption that the number of incidents in the data from JCH was much larger than expected, and that the purpose of his analysis was to discover whether there was an association with any of the nurses who worked on the ward. He multiplied his p-value for the association with de Berk's shifts by 27—the number of nurses in that period who worked on the same ward.

By the time he looked at the data from RCH, de Berk was a prime suspect and he judged that no further Bonferroni type correction was required. Finally, he proposed to take a very small probability for the significance level of his test.

In fact, his starting assumption was false: In the previous year, there had been no incidents in the ward, but the year before that, an even larger number. The hospital director had not revealed the information from two years ago to

the investigators, since the ward previously had a different name (he had changed it).

One could try to use a Bayesian approach to deal with the post-hoc problem. There would be good arguments for a rather low prior probability of an arbitrary nurse being a serial killer. The difficult task for the Bayesian approach would be determining a reasonable model for number of incidents if de Berk is a murderer, since one has to take into account that some proportion of the incidents are not murders at all.

Heterogeneity would also remain an issue for a Bayesian analysis. Explaining the methodology in a court of law could well be the biggest barrier.

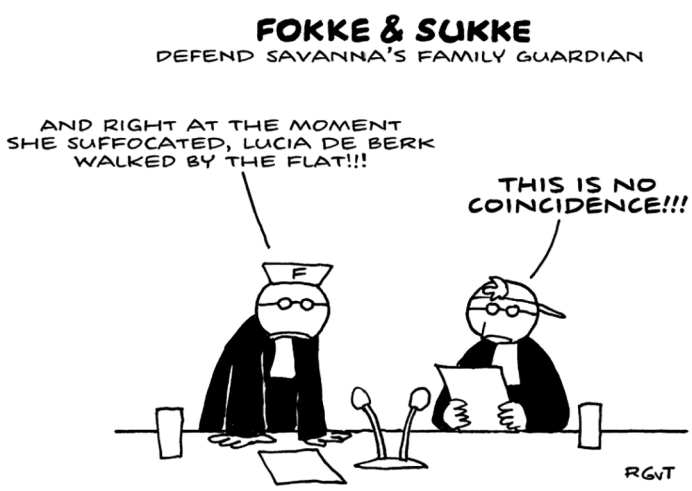


Figure 1. A Fokke & Sukke cartoon from October 30, 2007, in the Dutch newspaper *NRC Next*. (Translated into English by the creators of the cartoon, Reid, Geleijnse, and Van Tol). Lucia de Berk was still in prison at that time. The two ducks are defending a family guardian, accused of being responsible for the death of the girl Savanna, who died by suffocation. The accused woman was, in fact, acquitted (by another defense). What counselor Sukke is saying corresponds to what Elffers told the court: "Honored court, this is no coincidence. The rest is up to you."

An Alternative Model

The incidents that a nurse experiences can be modeled by a so-called Poisson process, with a nurse-dependent intensity A, using A for "accident proneness." A Poisson process is used to model incoming phone calls during non-busy hours, fires in a big city, etc. Since we believe the incidents to be rare, a Poisson process is an obvious choice for modeling the incidents that a nurse experiences.

This approach models two separate phenomena. Firstly, the intensity of nurses seeing or reporting incidents is modeled by introducing the random variable A. We assume that A has an exponential distribution, but other choices are also possible.

Note that we move away here from a simple discrete model, as used by Elffers, but use instead a continuous distribution for the "accident proneness" A of the nurse. Statistical models with continuously varying random variables might be more difficult to explain to the judges, but are often much more realistic, which should be the only important consideration here.

Secondly, the number of incidents happening to a nurse on duty depends on A and the time interval she is working, and follows (conditionally on A) a Poisson distribution. The time interval is measured by the number of shifts the nurse has had. Assuming that A is exponentially distributed implies, among other things, it can easily happen that one nurse has twice the incident rate of another nurse.

The probability of this event is 2 / 3; in fact the probability of an incidence rate of a factor k times that of another nurse is $2 / (k+1)$. The statistical problem boils down to the estimation of the parameter, characterizing the mixture of Poisson processes for the different nurses. Combining the Juliana Children's Hospital and the two wards of the Red Cross Hospital, de Berk had 201 shifts and 14 incidents.

A major flaw in the investigation is that the data collection is irreproducible and lacks rigorous methods and definitions. It crucially depended on the memory of people who knew what was sought after. We will argue from the data in Table 1, however, which was also used in Elffers's computations.

If the overall probability of an incident per shift is the ratio of total number of incidents to total number of shifts, $\mu = 27 / 1{,}734$. If a shift is the unit time interval, then this would be a so-called moment estimate of the mean intensity of incidents.

This means, that, conditionally on the time interval $T = 201$, the number of incidents follows a mixture of Poisson random variables with parameter $201A$, where the intensity A has an exponential distribution with first moment μ.

Thus, on average, an innocent de Berk would experience $201 \cdot \mu = 201 \cdot 27 / 1{,}734 \approx 3.12966$ incidents. A picture of

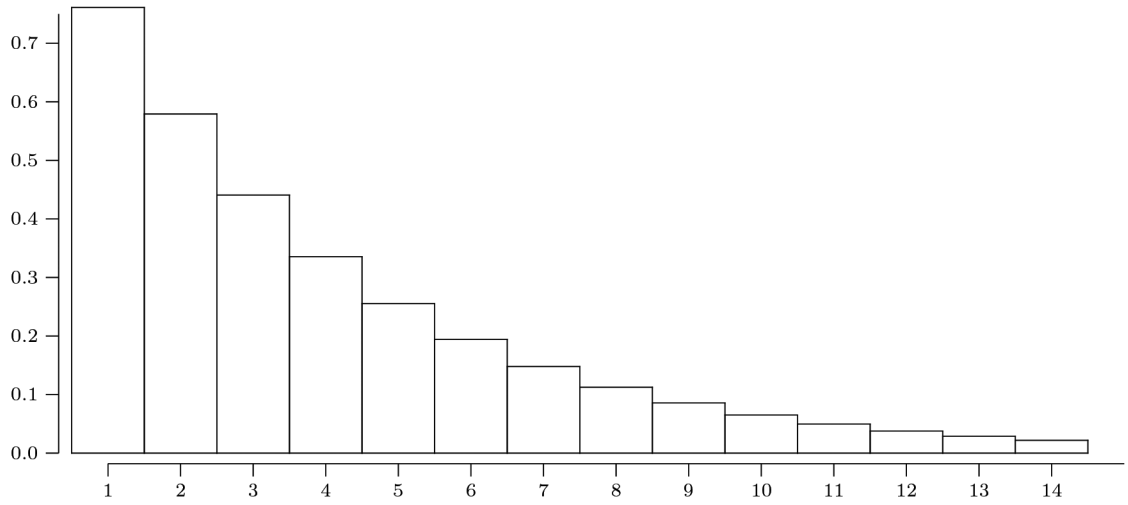


Figure 2. Probabilities (in the Poisson model) that the number of incidents in 201 shifts for one nurse is at least 1, 2, 3, … , if μ = 27 / 1,734. The probabilities are given by the heights of the columns above: 1, 2, 3…, respectively.

the probabilities that the number of incidents is greater than or equal to $k = 1, 2, \ldots$ is shown in Figure 2, which is based on the calculations in the next section. Heterogeneity of any kind increases the variation in the number of incidents experienced by a randomly chosen nurse over a given period of time (given number of shifts).

From the well-known relations for conditional expectation (E) and variances (var)

$$\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X|Y)),$$

$$\mathrm{var}(X) = \mathrm{E}(\mathrm{var}(X|Y)) + \mathrm{var}(\mathrm{E}(X|Y)),$$

it follows that whereas for a Poisson distributed random variable variance and mean are equal, for a mixture of Poisson's (with different conditional means), the variance is larger than the mean. If some nurses experience more or less incidents than others, the end result in all cases is overdispersion caused by heterogeneity.

Applied to the current model, which is geometric with parameter $(1 + t\mu)^{-1}$ (see the computation in the next section):

$$\mathrm{var}(N) = (1 + t\mu)^2 - (1 + t\mu) = t\mu + (t\mu)^2;$$

where the latter term neatly splits over the expected variance of the Poisson process plus the variance of the conditional parameter of the Poisson process, which is assumed to be exponential.

The fact that a modest amount of heterogeneity turns an almost-impossible occurrence into something merely mildly unusual is strong support for further empirical research on whether and, if so, in what forms, heterogeneity plays a role in healthcare. It can have major implications in different areas, such as medical research (representing an extra source of variation) and training medical staff.

Computation of the Probabilities in the Mixed Poisson Model

If N is a Poisson random variable with parameter λ, the probability that the number of incidents is bigger than k, $k = 1, 2, \ldots$, is given by an integral, namely

$$\frac{1}{(k-1)!}\int_0^{\lambda} e^{-x}x^{k-1}\,dx,$$

(see, e.g., Feller, Exercise 46, p. 173.)

This means that if we assume that the accident proneness of the nurses has an exponential distribution with expectation μ (in our case, estimated by 27 / 1,734) and the parameter of the Poisson distribution for the nurse is given by ta, where t is the time interval (in our case, t = 201), and a is the accident proneness, we have to integrate with respect to the density of the exponential distribution with expectation μ, taking $\lambda = ta$. For the probability that a nurse experiences or more incidents, that results in:

$$\int_0^{\infty} \mathbb{P}\{I \geq k | A = a, T = t\}\frac{e^{-a/\mu}}{\mu}\,da = \int_0^{\infty}\left\{\frac{1}{(k-1)!}\int_0^{ta} e^{-x}x^{k-1}\,dx\right\}\frac{e^{-a/\mu}}{\mu}\,da = \left(\frac{t\mu}{1+t\mu}\right)^k.$$

This is the geometric distribution with parameter $1/(1 + t\mu)$. With $k = 14$ and $t\mu = 3.18486$, this yields 0.020616 or about 1 in 49.

An early version of this paper used a revision of Elffers's data set proposed by Professor Ton Derksen, a philosopher of science, who with his sister, Metta de Noo, MD, was the first to actively contest the court's reasoning in the case of Lucia de Berk.

Our model led us to a right tail probability of 1 in 9. We later noticed that Derksen had also removed all incidents that the court finally decided not to count as provably caused by de Berk; he used the legal argument that Elffers had previously been instructed by the judges to do the same for the data from the Juliana Children's Hospital. This does not make any statistical sense.

Going back to original medical records, Derksen and de Noo also found inconsistencies in the classification and timing of several incidents, which underlines the unreliability of the data. Correcting the data for apparent errors would also improve the results from the defense point of view.

We decided to stick with Elffers's numbers here to focus on our main point concerning the impact of heterogeneity.

Extended Discussion of Heterogeneity

As shown, a modest amount of heterogeneity leads to very different orders of magnitude in the outcomes of crucial calculations. Some underlying mechanisms may lead to the postulated heterogeneity.

Clearly, the data in this case show heterogeneity. The data stem from two hospitals with very different patients: young children in the JCH and older adult patients in the RCH. The data come from three wards, and the rates of incidents per shift vary considerably for each ward.

Two general mechanisms cause heterogeneity. The first one concerns properties of subjects directly related to the intensity of the rate of incidents. The other mechanism is more indirect and results from "spurious correlations," in which properties that are not related to the underlying intensity influence the measurement via unexpected dependencies and systematic variations in variables assumed to be independent and uniform.

Related to this is another aspect of the data: the degree to which a specific model or null-hypothesis is susceptible to small variations in the data. This was the case in the original calculations. Although our example is tuned to this very specific case, it refers to a much more general caveat. It should be established how stable certain models are under small perturbations of the data.

Are Nurses Interchangeable?

According to medical specialists we have spoken to, nurses are completely interchangeable with respect to the occurrence of medical emergencies among their patients. However, according to nursing staff we have consulted, this is not the case at all. Different nurses have different styles and different personalities, and this can and does have a medical impact on the state of their patients. Especially regarding care of the dying, it is folk knowledge that terminally ill people tend to die preferentially on the shifts of those nurses with whom they feel more comfortable, although as far as we know, there has been no statistical research on this phenomenon.

There is another obvious way in which the intensity of incidents depends on characteristics that vary over the population. Any event that can turn out to be an "incident" starts with a call to a doctor, and in all cases, it is the nurse who decides to call a doctor. This decision is influenced by professional and personal attitude, past experience, and personality traits such as self-confidence. It seems obvious that these characteristics vary greatly in any population, so we assume that the intensity of experiencing incidents varies accordingly.

Inadequacy of the Hypergeometric Distribution as a Model and Spurious Correlations

The model underlying the null-hypothesis (which led to the hypergeometric distribution) depends on two assumptions: Both the incidents and the nurses are assigned to shifts uniformly and independently of each other.

We have established two ways in which characteristics of individual subjects may lead to variation in the intensity of experiencing an incident. This variation is in contrast with one of the assumptions underlying the hypergeometric distribution: uniformity.

Sources of correlation can correspond to indirect rather than direct causation; spurious correlation can be explained by confounding factors and common causes.

There are serious reasons to doubt the uniformity of incidents over shifts. Periodical differences may occur. The population of a hospital ward may vary over the seasons. The patients may differ in character and severity of illness, due to seasonal influences. There are differences between day and

night shifts, and between weekend shifts and shifts on weekdays. An extensive study of Dutch intensive care units admissions shows a marked increase in deaths when the admission falls outside "office hours" (see Hans, Kuijsten, Brinkman, Meynaar, et al.)—there have to be nurses on duty throughout the night and throughout the weekends, while the medical specialists tend to have "normal" working hours.

Finally, there is the periodical cycle of the circadian rhythm, influencing the condition of the patients and the attention of the medical staff (see Kuhn).

Notice that circadian variation in areas such as mortality and the resulting variation of incident rate between different shifts over the day interacts with the variation in the number of nurses on a shift, with more personnel on the day shifts. This can result in a higher number of nurses with an incident on their shift if the incident rate is higher during daytime shifts and, conversely, a lower number in the opposite case.

There may be other, nonperiodical variations that affect the uniformity of incidents. In the case of the Juliana Children's Hospital, there has been a rather sensitive matter of policy: whether very ill children, who are not going to live for very long, should die at home or in hospital wards. We understand that this policy did change at least once at the JCH in the period of interest. Presumably, a change in policy concerning where the hospital wants children to die, will have an impact on the rate of incidents. Further, incidents may be clustered, since one patient can give rise to several incidents.

On the other hand, the way nurses are assigned to shifts is certainly not uniform and "random". Nurses take shifts in patterns; for example, several night shifts in a row, alternated by rows of evening or day shifts. Nurses are assigned to shifts according to skills, qualifications, and other characteristics. Some nurses might take relatively more weekend shifts than others because of personal circumstances.

Although the assignment of both nurses to shifts and incidents to shifts are not uniform processes, one could hope that there might be some "mixing" condition that makes the ultimate result indistinguishable from the postulated independence and uniformity. This magical mechanism should at least be made plausible.

Taken together, even if we consider both the shifts of a given nurse and the incidents on a ward as random processes, and the two processes as stochastically independent of one another, the assumption of constant intensities of either is a guess, not based on any evidence or argument.

There may be patterns in the risk of incidents, and there are certainly patterns in the shifts of nurses. These patterns may be correlated, through the process by which shifts are shared over the different nurses according to their different personal situations, their different wishes for particular kinds of shifts, their different qualifications, and the changing situation on the ward.

How Stable are the Hypergeometric Probabilities With Small Changes in the Data?

Consider the data for the ward at JCH. These numbers and their interpretation are at the root of what turned out to be one of the gravest miscarriages of justice in the Dutch judicial system. Under the assumption of the hypergeometric distribution, the probability of this configuration is very small; less than 1 in 9 million.

The configuration is, in some respects, extreme: Eight out of eight incidents occur during the shift of one nurse. However, the data are, in another respect, also conspicuous: no incidents occur in the 887 shifts where this nurse was not present (see Table 1). The data collection had been far from flawless, with no formal definition of incident, and no or incomplete documentation, and rested at least in part on recollection of witnesses who were aware of which facts were looked for.

Assuming the possibility of tiny flaws in the process of data acquisition, it is legitimate to investigate the effect of $1, 2, \ldots, 8$ incidents that could have been forgotten or overlooked. This amounts to allowing a maximal error of less than 1 percent. The results are quite remarkable; Table 2 shows the probabilities.

The very small numbers vanish easily. Six or more incidents not remembered, not reported, or just defined away make the difference between astronomically small on the one hand and very unusual on the other. This shows that the probabilities are quite sensitive to small errors in the data.

A judgment on data quality is not only the concern of a statistician. Judges are used to inconsistent and incomplete data (statements); psychologists are very well aware of the possible fallacies of memory. Both groups have their own professional standards for how to deal with these phenomena. A statistician, however, should point out what the effects of these phenomena can be on the outcome of his models.

If this model is used to corroborate evidence, this sensitivity should be made explicit, just as adverse workings of a medicine are mentioned explicitly for the users.

Table 2—The Effect of Perturbations on the Probabilities

Shifts with incidents outside Lucia de Berk's (postulated)	0	1	2	3	4
Probability	1/9,043,864	1/1,137,586	1/257,538	1/79,497	1/29,989
Shifts (continued)	**5**	**6**	**7**	**8**	
Probability	1/13,051	1/6,329	1/3,341	1/1,889	

Concluding Remarks

We have shown the considerable effect that a modest amount of heterogeneity can have on tail probabilities. The broader impact of allowing heterogeneity in the analysis of (medical) research has interesting consequences beyond the case of Lucia de Berk. What remains is a very short description of how the case ended in acquittal.

de Berk was arrested in December 2001. The court (of appeals) stated that it did not include statistical considerations as a basis for its verdict. This may be true for formal statistical considerations, but the essential step in the construction of the guilty verdict was that only one or two cases of murder had to be proven convincingly; the rest of the murders could be considered proven based on the "very improbable" occurrence of incidents during the shifts when de Berk was working. In this way, statistical considerations were crucial, but the verdict was immunized against formal statistics. de Berk was convicted in 2004 for seven murders and three attempts of murder. What followed was a long legal struggle where the emphasis was on the validity of the medical arguments and increasingly intricate juridical matters.

The de Berk case was fiercely debated in public and the statistical notions remained an important issue. Statisticians, now banned from the courtrooms, continued to play a role; for example, by mobilizing the scientific community. Gradually, the notion emerged that a gross miscarriage of justice had taken place. A complicating factor remained that, since the judicial path had been followed until the end, a new "fact"—a so-called *novum*—had to be found.

In 2008, de Berk was allowed to wait for the end of the legal proceedings from outside prison, and two years later, she was finally acquitted of all murder accusations. C

Further Reading

Elffers, H. May 8, 2002. Distribution of incidents of resuscitation and death in the Juliana Kinderziekenhuis [Juliana Children's Hospital] and the Rode Kruisziekenhuis [Red Cross Hospital]. Unpublished report to the court. *http://www.math.leidenuniv.nl/~gill/Elffers1eng.pdf*.

Elffers, H. May 29, 2002. Distribution of incidents of resuscitation and death in the Juliana Kinderziekenhuis [Juliana Children's Hospital] and the Rode Kruisziekenhuis [Red Cross Hospital]. Unpublished report to the court. *http://www.math.leidenuniv.nl/~gill/Elffers2eng.pdf*.

Feller, William. 1968. *An introduction to probability theory and its applications*, Vol. 1, Third edition. New York-London-Sydney: John Wiley & Sons, Inc.

Hans, A.J., Kuijsten, M., Brinkman, Sylvia, Meynaar, Iwan A., et al. 2010. Hospital mortality is associated with ICU admission time. *Intensive Care Medicine*, doi: *10.1007/s00134-010-1918-1*.

Kuhn, G. 2000. Circadian rhythm, shift work, and emergency medicine. *Annals of Emergency Medicine*, *37*:88-98. doi: *10.1067/mem.2001.111571*.

Meester, R., Collins, M., Gill, R.D., and M. van Lambalgen. 2007. On the (ab)use of statistics in the legal case against the nurse Lucia de B. *Probability and Risk*, *5*:233-250. With discussion by David Lucy.

Schneps, L., and Colmez, C. 2013. *Math on trial. How numbers get used and abused in the courtroom*. New York: Basic Books.

About the Authors

Richard Gill is emeritus professor of mathematical statistics at Leiden University. He is interested in statistics and the law, as well as statistics in quantum physics, survival analysis, and semiparametric models.

Piet Groeneboom is emeritus professor of statistics of Delft University of Technology. His present research focuses on shape-restricted inference and computer algorithms for non- and semiparametric statistical models.

After finishing his PhD thesis on central limit theorems with applications in Malliavin calculus, **Peter de Jong** started his company on internet publishing. He taught statistics for actuaries. He is also editorial advisor to the scientific publisher Ios Press on the role of statistics in their portfolio.

A Simple Example of Isolation in Building a Natural Experiment

Jose R. Zubizarreta, Dylan S. Small, and Paul R. Rosenbaum

What is isolation?

Large data systems increasingly record outcomes and exposures to treatments, but often the occurrence of an exposure/outcome event, the timing of the event, physical location of the event, and attending circumstances are biased in ways that cannot be controlled using data recorded by that system. In this example of isolation (an especially simple example) the data are from the U.S. Fatality Analysis Reporting System (FARS), which records information about motor vehicle accidents with at least one fatality. The system records interesting information about such accidents, including safety equipment, safety belt use, airbag deployment, and much more.

Arriving after the fact, the system does not record reliable information about speeds, the initial distance between vehicles before one vehicle started to brake, efforts the drivers made or failed to make to limit the extent of a collision, the promptness and effectiveness of braking, even though these are important determinants of the forces involved in the crash.

In a crash, as in any physical event, forces matter. If Volvo station wagons are driven more safely than Dodge Challengers, at slower speeds, with fewer high-speed lane changes, and at a greater distance from the car ahead, none of this is recorded in FARS, so superior safety equipment in the Volvo may be credited with effects actually produced by safer driving. If convertible coupes are more often driven in weather with excellent road traction, while four-wheel-drive SUVs are driven with less attention to weather, then convertibles may appear safer than they would be if driven with poor traction and, hence, worse braking.

Importantly, FARS records nothing about the crash that did not occur because of anti-lock brakes, nor the crash in which all fatalities were prevented by a treatment under study, such as safety belts and airbags. Every crash in FARS in which a lone driver hit a tree is a lethal crash, not because trees are so dangerous, nor because safety equipment is helpless against trees, but because, in these circumstances, only a lethal crash would be recorded in FARS.

In FARS, there are many obstacles to judging the crashworthiness of a Volvo station wagon or a Ford F450 pickup truck, because it is difficult to separate the vehicles from the way people drive them; however, these are not insurmountable obstacles.

Within a large but biased data system, isolation means constructing or finding a smaller set of comparisons that are likely to be much less biased.

In a biased data system, there may be a fairly large and not very biased natural experiment. One isolated comparison that Evans (1986) and Rosenbaum (2015) found in FARS, but was not constructed, compares a driver and a right-front passenger in the same car; one wearing a safety belt, the other unbelted. That assesses the effects of safety belts while controlling unmeasured speeds, braking, road traction, and some consequent forces.

Norvell and Cummings (2002) made another isolated comparison, assessing the effects of motorcycle helmets by comparing two riders of the same motorcycle; one with a helmet, the other without.

Isolation means that the occurrence of a treatment/outcome event, its timing, its location, and its circumstances may be severely biased, but conditionally that one treatment is in lieu of another at that time, in that place, and under those circumstances, the conditional distribution of exposure to these two treatments is either not biased or much less biased. According to Zubizarreta, et al., this can be shown formally: Certain types of unmeasured biases in the hazard of treatment drop out when conditioning on receiving one treatment in lieu of another.

Found comparisons are immediately easy to appreciate, but the possibility of optimal construction substantially enlarges the scope of isolation. For example, the effect of an additional child on a woman's career is usually difficult to assess, because many women are active in controlling their fertility to fit their career plans. That control may be lost when a mother finds herself pregnant with twins; a study by Joshua Angrist and William Evans used this fact to obtain estimates of the effects of an additional child on a woman's career. Building upon their idea, the authors of this article matched mothers with the same past history of fertility and education at the moment that one mother had twins and her five matched controls had single births.

More precisely, isolation refers to a differential comparison in a risk-set match. A differential comparison compares one treatment given in lieu of another treatment. In an isolated comparison, a risk-set match compares people who received one treatment in lieu of another at one of those rare and brief moments when the biases that promote one treatment and not the other are absent.

Perhaps not wearing safety belts is more common among reckless drivers, but that matters much less, if at all, given two people in the front seat of the same car in the same crash with one belted, the other unbelted. Perhaps wearing a motorcycle helmet and riding cautiously typically go together, but that matters much less, if at all, given two people on the same motorcycle, one with a helmet, the other without. Perhaps a woman's plans for career and family strongly bias the occurrence and timing of births, but that matters much less, if at all, when comparing two pregnant women with the same history of fertility and education, one currently having twins, the other having a single child.

Outside randomized experiments, the effects caused by treatments are difficult to see—they are shrouded in fog. For brief moments, on rare occasions, the fog lifts and an equitable comparison emerges from the mist. Isolation refers to noticing and collecting those brief and rare occasions, and separating them from everything else.

Which Cars are Safe in Crashes

The Insurance Institute for Highway Safety (IIHS) conducts experimental crash tests in which cars strike fixed objects and reports results on its webpage. Their efforts are valuable, but they mention an important aspect but do not emphasize it. They wrote:

> "IIHS conducts vehicle tests to determine crashworthiness—how well a vehicle protects its occupants in a crash … Models that earn Top Safety Pick+ or Top Safety Pick are the best vehicle choices for safety within size categories. Size and weight influence occupant protection in serious crashes. Larger, heavier vehicles generally afford more protection than smaller, lighter ones. Thus, a small car that's a Top Safety Pick+ or Top Safety Pick doesn't necessarily afford more protection than a bigger car that doesn't earn the award."

IIHS often notes the role of vehicle mass, and then the page gives "Top Safety Pick" awards to two "minicars," a 2017 Mini Cooper and a 2017 Toyota Yaris, and to one "large pickup," the 2017 Honda Ridgeline. Other types of cars, minivans, and SUVs also won awards. The Ford F-450 pickup won no award for safety, but has a shipping weight that can be more than three times the weight of a Mini Cooper. How much does that matter?

A large biased data system may contain within it a smaller, much less-biased natural experiment. Isolation finds or constructs the natural experiment.

Is safety in an automotive category an important indicator when vehicles in different categories collide? Which vehicles are safer in accidents involving two vehicles?

There are many interesting questions about automotive safety, and FARS can speak to some of them. This particular question is not about what causes vehicles to collide, nor about what causes them to collide with particular forces; rather, it asks, What happens conditionally given that two specific vehicles from different categories did collide with the forces that were actually present? In other words, What is the differential effect of being in one type of vehicle in lieu of another type, given an accident involving these two vehicles in particular, largely shared, circumstances?

Many things that might bias a comparison between different accidents are made similar by

comparing two vehicles in the same accident.

> What is the differential effect of being in one type of vehicle in lieu of another type, given an accident involving these two vehicles in particular, largely shared, circumstances?

Perhaps convertible coupes are more often driven on sunny days with excellent road traction, while four-wheel drive SUVs are often driven in snow, so a collision between a convertible and an SUV is likely to occur with better braking than many collisions between two SUVs. However, given the condition that a convertible and an SUV have collided, with the forces that did exist in that crash, what is the differential effect of being in the convertible rather than the SUV?

The variation in forces between different collisions at different times and places may be biased by weather conditions and driving behaviors that increase or reduce the hazard of certain types of accidents. However, in any collision, the opposing forces are dissipated, perhaps in a complex way, in bringing the two colliding vehicles to a halt.

Think of adding opposed force vectors in high school physics. Perhaps Volvo station wagons are typically driven more safely than pickups, at slower speeds with greater distance from the car ahead, so the typical accident involving a Volvo wagon might be a lower-speed, lower-force event than the typical accident involving a pickup. However, this matters much less, if at all, conditionally, given that the Volvo and the pickup have collided in a particular way. What is the differential effect of being in the Volvo wagon rather than the pickup in this accident as it actually happened?

> At any moment while driving, everyone has a momentary risk or hazard of being involved in a lethal crash—a risk that varies with time, weather, driving behavior, and many other factors. As a safe driver, your hazard of being in a lethal crash may be lower than mine, and you may be more likely to be in a slower-speed, lower-force crash, but all that matters less, if at all, conditionally given that we have collided in a single crash. Before the fact, our hazards of being in a crash of this kind may have been very different, but all this matters less conditionally, given that this crash has occurred.

Is there evidence of bias in which crashes occur under what circumstances? Is there evidence that it is unwise to look at vehicles one at a time? Indeed there is.

The biases we can see are less worrisome than the biases we cannot see, because we can often correct for biases we can see, but the biases we can see suggest such biases are commonplace.

In 2,923 crashes involving two moving vehicles, one a car, the other a light truck, 64 crashes involved a convertible car, but only four crashes involved a convertible car in the winter in the northern U.S.—specifically, in November, December, January, and February, at or above latitude 40 degrees, or north of Philadelphia. In these same 2,923 crashes, pickups were 75% more likely than SUVs to be in a fatal crash on a country road (an odds ratio of 1.75), presumably because of where pickups tend to be driven, not because of their unusually poor performance on country roads.

In contrast, in a single crash involving two vehicles, many things, measured or unmeasured, are typically the same or similar for the two vehicles: ambient light, weather, location, road traction, inter-vehicle distance before braking, and much more.

> Because different types of vehicles, say a Volvo station wagon and a Ford F-450 pickup, are driven by different people, in different ways, under different circumstances, the type of vehicle may predict both the absolute risk of a crash and unmeasured aspects of a crash, such as the distance between vehicles before the start of any braking and the reduction in relative speed due to braking. However, given that a Volvo station wagon and a Ford F-450 pickup have collided, the type of vehicle predicts fewer features of the crash; for instance, the distance between the two vehicles before any braking is the same.

Isolated Comparisons of Crash Safety

Description of the comparison

Using data from the 2015 FARS, we examine lethal accidents involving exactly one car and one so-called "light truck." A light truck is an SUV, minivan, or pickup. Each individual accident—each such isolated comparison—produces its own 2 x 2 table of counts, car or light truck, survived or died. There are 2,923 such

Table 1—Five of the 2,923 Tables Recording Mortality in Crashes of One Car and One Light Truck

ID	Type	Mortality		Total	Make/Model	Weight (lbs)
		Survived	Died			
10060	Light Truck	2	0	2	Honda Odyssey	4,365
	Auto	4	1	5	Ford Escort	2,419
	Total	6	1	7		
		Survived	Died			
10114	Light Truck	0	1	1	Ford Explorer	3,952
	Auto	1	0	1	Acura TL	3,623
	Total	1	1	2		
		Survived	Died			
10129	Light Truck	1	0	1	Nissan Pathfinder	3,961
	Auto	1	1	2	Pontiac Gand Am	3,102
	Total	2	1	3		
		Survived	Died			
10130	Light Truck	2	0	2	GMC Yukon	4,947
	Auto	0	1	1	Pontiac Bonneville	3,633
	Total	2	1	3		
		Survived	Died			
10195	Light Truck	1	0	1	GMC New Sierra	5,284
	Auto	0	2	2	Ford Fusion	3,526
	Total	1	2	3		

Weight refers to shipping weight; that is, the weight of the vehicle without passengers or cargo. The ID number from FARS is STCASE.

2 x 2 tables; Table l displays five of the 2 x 2 tables.

For both light trucks and cars in this comparison, the minimum and median number of occupants was l and the upper quartile was 2, but the maximum for light trucks was l2 and for cars was 7.

A few details of the comparison should be noted. This figure/study considers only accidents in which there are exactly two vehicles (VE_TOTAL==2), both of which are in transit—e.g., not parked—(VE_FORMS==2), and the only people involved in the crash were in one of these two vehicles—no pedestrians or bicyclists—(PERNOT - MVIT==0).

There had to be data in FARS about both vehicles. A few two-vehicle accidents were not collisions—perhaps a vehicle went off the road to avoid a collision—but were not removed from the analysis.

The vehicles had to be either passenger cars or trucks (VEHTYPE_T=="Passenger Car" or VEHTYPE_T=="Truck"), excluding, for instance, motorcycles.

Finally, attention was restricted to cars and light trucks by requiring BODYSTYL_T to be one of the following categories: "Convertible," "Coupe," "Hatchback," "Pickup," "Sedan," "Sport Utility Vehicle," "Van Cargo," "Van Passenger," or "Wagon." Light trucks were defined to have BODYSTYL_T of "Pickup," "Sport Utility Vehicle," "Van Cargo," or "Van Passenger," the rest being cars.

By definition, every 2 x 2 table contains at least one person in a light truck, at least one person in a car, and at least one death, but there were no survivors in 68 of the 2,923 accidents.

The comparisons illustrate the statistical concept of isolation,

but do not endorse or critique particular vehicles or brands, for which numerous considerations are relevant.

Does it Matter that FARS Only Records Lethal Crashes?

An accident is in FARS only if someone dies; the FARS system attempts to describe every crash in which someone died. Suppose for this discussion that it was successful in that attempt. How does the $2 \times 2 \times 2{,}923$ table from FARS relate to the much larger $2 \times 2 \times K$ table that records all accidents involving one car and one light truck in 2015? The larger table has $K - 2{,}923$ subtables in which no one died. That is, the 2×2 tables not in FARS differ from those in Table 1 in that each column total in the second—or "Died"—column is zero.

The best test of no association between vehicle type and mortality within crashes—more precisely the uniformly most powerful unbiased test against a constant odds ratio not equal to 1 in a $2 \times 2 \times K$ table—conditions on the four row and column totals in each 2×2 subtable, the so-called marginal totals. The familiar Mantel-Haenszel test is one of several large sample approximations to this best, exact test (Birch, 1964).

In both the exact test and its large sample approximation, a table with a zero in the row or column totals does not contribute to the test. That is because if there is a zero in the margin, then there is only one 2×2 table with the same margins as this table. In this case, conditioning on the margins has fixed the entire table, so it is constant and not random. Intuitively, a 2×2 table in which no one died may be a crash were slow speeds and successful braking limited the forces involved in the crash, and such a crash provides little or no information about crashworthiness and mortality.

In parallel, tables with a zero marginal total do not affect the conditional maximum likelihood estimate of a constant odds ratio or of coefficients describing the dependence of odds ratios on covariates. The conditioning in the $2 \times 2 \times K$ tables is the simplest case of the more-general type of conditioning in Zubizarreta, et al. (2014, §2.3).

In brief, FARS provides all of the 2,923 tables with at least one death and omits the $K - 2{,}923$ tables without a death, but the omitted tables would not have affected the conditional analyses, even had they been present. The FARS system is not a census of U.S. vehicles in 2015—presumably, vehicles driven by cautious drivers are underrepresented—nor is it a census of vehicle crashes in 2015—presumably, less-forceful crashes are underrepresented. However, the $2 \times 2 \times 2{,}923$ table is essentially a census of lethal crashes involving exactly one car and one light truck, so it is a census of the informative part of the $2 \times 2 \times K$ table recording all such crashes.

Cars versus Light Trucks

When a car and a light truck collide, who dies? The answer can be expressed in terms of odds ratios. If you flip a fair coin and roll a fair die, then the odds of a "head" on the coin are 1-to-1 or 1/1, and the odds of "6" on the die are 1-to-5 or 1/5, so the odds ratio is (1/1) / (1/5) = 5.

This means that the head is five times more likely than the 6; its odds are five times greater. (We calculate odds ratios within crashes using the clogit function in the survival package in R, and we report 95% confidence intervals, CI, for odds ratios within parentheses.)

In the $2 \times 2 \times 2{,}923$ table of 2,923 lethal crashes of a car and a light truck, the estimated odds ratio linking occupancy in the car with death is 4.76 (with 95% CI [4.34, 5.23]). In brief, the odds of death are almost five times higher in the car than in the light truck.

Safety belts also matter. Four types of restraints can be distinguished: lap/shoulder belts (71.7%), child restraints (3.0%), no belt restraint (21.5%), and other (3.9%). Being unrestrained rather than wearing a lap/shoulder belt was associated with a 6.19-fold higher odds of death. Taking both restraint use and vehicle type into account, being in a car rather than a light truck was even more strongly associated with death: The odds ratio is 5.35, rather than 4.76 (with 95% CI [4.79, 5.97]).

Of the 2923 lethal crashes involving a car and a light truck, there were 14 crashes in which the car was one of the two brands of minicars mentioned by the IIHS, the Mini Cooper or the Yaris, though the 14 cars were not the specific years and models IIHS recommended. There were 48 people involved in these 14 crashes, of whom 14 died. All 14 deaths occurred in the minicars, although eight occupants of minicars survived. Using the exact test of Birch (1964), the two-sided *P*-value testing no association within crashes is 0.0000643.

Lighter Cars versus Heavier Light Trucks

Some vehicles are labeled light trucks because of their body styles, but are not hefty vehicles. For instance, the Chrysler PT Cruiser is nominally a passenger van,

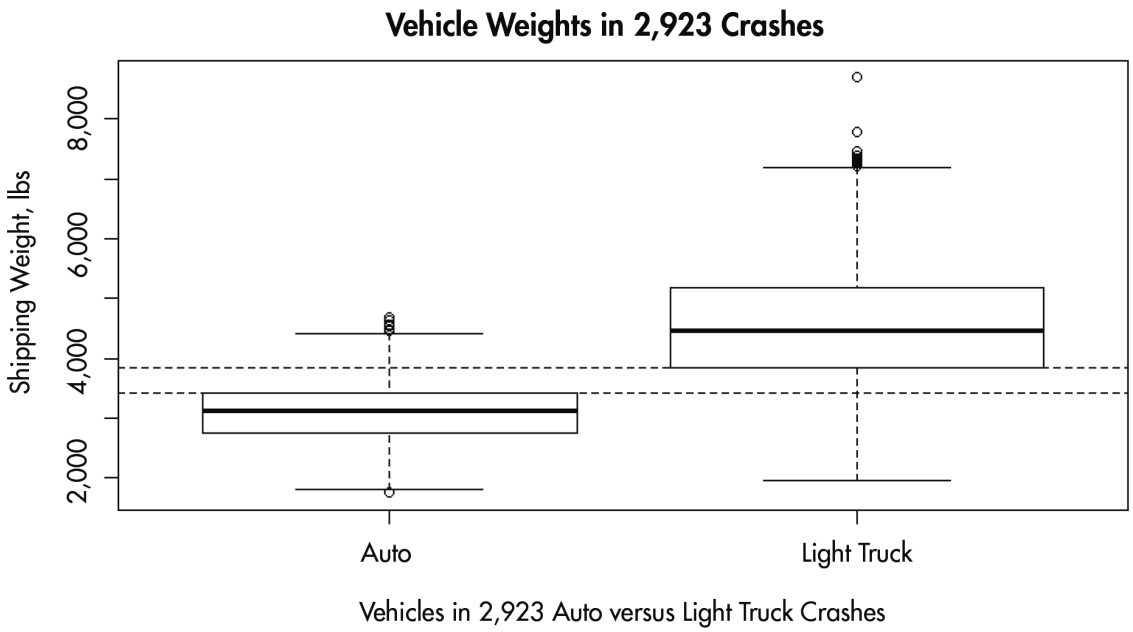


Figure 1. Vehicle weights in 2,923 lethal crashes involving exactly one auto and one light truck. The shipping weight is from the manufacturer, and does not include passengers and cargo. The two dashed lines are at the upper quartile of auto weights, 3,422 lbs., and the lower quartile of light truck weights, 3,847 lbs.

a type of light truck, but at 3,165 lbs, it weighs somewhat less than a Honda Accord; the same is true of some Ford Ranger pickups. Additionally, some of the larger cars from Lincoln and Mercedes-Benz weigh as much as the typical light truck—about 4,450 lbs. The comparison in Figure 1 is not quite the intended comparison of hefty SUVs with commuters' cars.

The 2,923 crashes in Figure 1 are divided roughly in half, into 1,565 crashes on left in Figure 2 (see p. 22), and 1,358 crashes on the right in Figure 2. Figure 2 labels these as typical or atypical, but it would be more correct to speak of prototypical and ambiguous. The division is based on the dashed lines in Figure 1, at the upper quartile for cars and the lower quartile for light trucks.

On the left in Figure 2 are 1,565 crashes of a car whose weight is below the upper quartile for cars with a light truck whose weight is above the lower quartile for light trucks.

On the left in Figure 2, the median car weights about 3,000 lbs., and the median light truck weighs about 4,750 lbs—about 50% more. On the right in Figure 2 are the remaining 1,358 crashes: a heavier car, a lighter light truck, or both, or a missing vehicle weight.

Although the situations on the left and right of Figure 2 are about equally common in FARS, the situation on the left is the intended or prototypical situation, in which a car that is not especially large collides with a light truck that is not especially small. A collision between a Lincoln Town Car and a Chrysler PT Cruiser would be on the right of Figure 2, and would roughly reverse the median weights on the left—3,000 versus 4,500. If you were trying to decide between a hefty SUV and a small or mid-sized car with safety in mind, then the left side of Figure 2 is more relevant than Figure 1.

For the comparison on the left in Figure 2, the odds ratio is 8.10 (with 95% CI [6.99, 9.39]), while on the right, it is 2.91 (with 95% CI [2.57, 3.30]). Taking safety belt use into account, as above, yields an odds ratio on the left of 9.16 (with 95% CI [7.70, 10.90]), while on the right, it yields an odds ratio of 3.24 (with 95% CI [2.80, 3.75]). When a hefty SUV, van, or pickup collides with a small or mid-sized car, the odds of death are about eight or nine times higher in a car.

Could it be that robust, survival-prone individuals buy pickups, and frail, mortality-prone individuals buy small and mid-sized cars, so

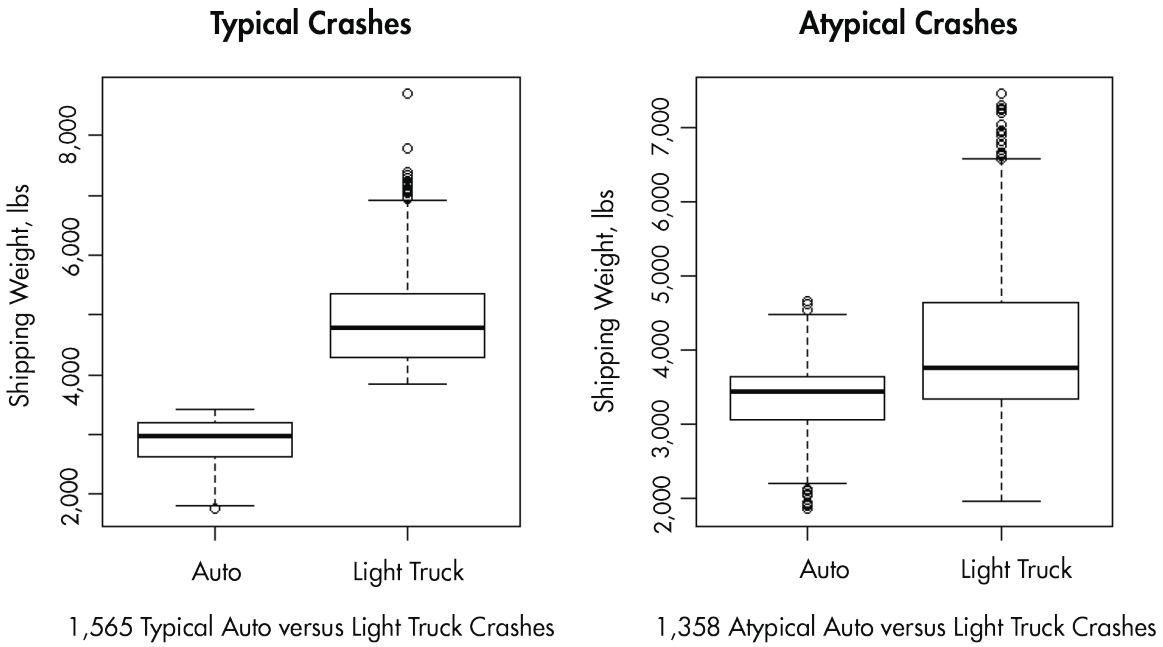


Figure 2. The 2,923 auto/light-truck crashes are divided into two groups of 1,565 typical crashes and 1,358 atypical crashes; 2,923 = 1,565 + 1,358. In typical crashes, the auto had a weight that was at most the upper quartile of auto weights, 3,422 lbs., and the light truck had a weight that was at least the lower quartile of light truck weights, 3,847 lbs.; see the two dashed lines in Figure 1. All other crashes are atypical: Either the auto was heavy or the light truck was light or both, or at least one weight was missing in 112 crashes.

the results above are entirely due to bias in who buys light trucks, not an effect caused by the differences in mass in Figure 2?

Actually, the comparison on the left in Figure 2 is insensitive to large unmeasured biases. It is easiest to think about a matched pair—a car and a light truck whose only occupants are their drivers. Of the 2,923 crashes, 1,129 were matched pairs. Consider the magnitude of unmeasured bias needed to explain away the survival difference between hefty light trucks and conventional cars on the left of Figure 2. In a matched pair, to explain away the survival difference, that bias would increase the odds of buying a light truck 13-fold and increase the odds of death 15-fold; see Rosenbaum (2017; Table 9.1, with $\Gamma = 7$).

To explain the difference in survival in terms of the unmeasured frailty of people who buy cars instead of pickups, there would have to be an enormous difference in frailty closely tracking who buys what.

Some Reminders, Considerations, and Limitations

Sometimes, the hazard of the timing and occurrence of a treatment is biased by unobserved covariates, but some of these biases may vanish briefly in a differential comparison of two treatments, conditionally given that a particular comparison has occurred at a particular time. This is isolation.

Cautiously driven cars may typically be involved in fewer, slower, less-forceful crashes, but this matters less conditionally given that two particular vehicles have collided. Isolation may occur naturally, scattered throughout an otherwise-biased data system, or isolation may be deliberately constructed by the analyst (see Further Reading for an example of a construction).

Isolation may remove some important unmeasured biases, but cannot be relied upon to remove all unmeasured biases. Differential biases may remain. Perhaps cautiously driven cars are typically struck, while recklessly driven cars typically strike, so isolated

comparisons may remain somewhat biased. For this reason, analytical adjustments and sensitivity analyses are needed to address any differential biases that may remain (see Further Reading).

Isolation attempts to find a natural experiment in a population, but it does not attempt to represent or describe the population as it naturally occurs. Experiments that estimate causal effects and surveys that faithfully describe populations are different things, and they can be biased in different ways.

Because certain types of cars may often be driven more cautiously than others, equitable comparisons of crash safety may occur only in atypical crashes; that is, crashes that are unrepresentative of the population. If Volvo station wagons are typically driven more safely than large pickups, then it might be wise to charge less to insure the station wagon, even though you might be safer in the large pickup, conditionally given a collision between these two vehicles.

If the population does not resemble an equitable experiment, then an equitable natural experiment will not resemble the population.

Conclusion

A large but biased data system may contain a smaller, much less-biased, natural experiment. Some isolated comparisons occur naturally and simply have to be noticed and separated from other, more numerous, but biased comparisons. Other isolated comparisons combine a random element that has to be noticed and a construction that pulls together similar situations in which that random element picked one treatment in lieu of another.

Further Reading

Angrist, J.D., and Evans, W.N. 1998. Children and their parents' labor supply. *American Economic Review, 88*, 450–477.

Birch, M.W. 1964. The detection of partial association, I: The 2 x 2 case. *Journal of the Royal Statistical Society*, B, 313–324.

Evans, L. 1986. The effectiveness of safety belts in preventing fatalities. *Accident Analysis and Prevention, l8*, 229–241.

Insurance Institute for Highway Safety. 2017. *http://www.iihs.org/iihs/ratings/TSP-List*. Accessed August 5, 2017.

National Highway Traffic Safety Administration. 2015. Fatality Analysis Reporting System. *www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars*.

Norvell, D.C., and Cummings, P. 2002. Association of helmet use with death in motorcycle crashes: a matched-pair cohort study. *American Journal of Epidemiology, l56*(5), 483–487.

Rosenbaum, P.R. 2006. Differential effects and generic biases in observational studies. *Biometrika*, *93*, 573–586.

Rosenbaum, P.R. 2015. Some counterclaims undermine themselves in observational studies. J*ournal of the American Statistical Association, ll0*, 1,389–1,398.

Rosenbaum, P.R. 2017. *Observation and Experiment: An Introduction to Causal Inference*, Chapter 12. Cambridge, MA: Harvard University Press.

Zubizarreta, J.R., Small, D.S., and Rosenbaum, P.R. 2014. Isolation in the construction of natural experiments. *Annals of Applied Statistics, 8*, 2,096–2,121.

About the Authors

José R. Zubizarreta is an assistant professor of statistics in the Department of Health Care Policy of the Harvard Medical School.

Dylan S. Small and **Paul R. Rosenbaum** are professors of statistics at the Wharton School of the University of Pennsylvania.

"With Great Power Comes Great Responsibility": Harnessing the Extraordinary Power of Statistics

Daniel H. Mowrey, Jonathan Potts, Susan Spruill, Walter Stroup, and Michiko I. Wolcott

The call for banning *p*-values by the editors of a noted scientific journal in 2015 continues to frustrate statisticians. While such controversies typically surface around a specific statistical concept such as *p*-values, they are symptoms of a larger problem that includes communication gaps and mistrust of statistics in the scientific/business communities.

Ironically, the call for the *p*-value ban was motivated by the widespread misuse of the *p*-values in the name of science. Statisticians have an obligation to the discipline of statistics and to the fields of application to promote statistical best practices while discouraging misuse and abuse that come from poor understanding of statistical tools.

At the Joint Statistical Meetings in Chicago in August 2016, a group of panelists from varying statistical backgrounds—Daniel Mowrey (DM), Jonathan Potts (JP), Susan Spruill (SS), and Walter Stroup (WS)—moderated by Michiko Wolcott (MW), deliberated the concept that the *p*-value is not the problem, but rather a symptom of other problems, namely the lack of early involvement by the statisticians as collaborators and by gaps in education on the importance and utility of correctly interpreting statistical analyses.

The following is an interview with the panelists.

MW: What are the key perceptions about statistics and statisticians in the world around us?

DM: To the non-statistician, statistics goes hand in hand with sports or anything associated with numbers. The person who is responsible for summarizing these numbers is, of course, the statistician.

People tend to fall into two camps with respect to statistics: 1) Statistics can be useful in proving a point or 2) Statistics are almost the lowest thing on the earth. The expression "lies, damned lies, and statistics" summarizes these two points quite nicely.

The scientific community probably views our discipline similarly to the Merriam-Webster definition: "A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data." The American Statistical Association (ASA) describes statistics as the science of learning from data, and of measuring, controlling, and communicating uncertainty; thereby, providing the navigation essential for controlling the course of scientific and social advances.

There are many interpretations associated with statistics. However, I believe statisticians would like to convey to everyone that it is a fascinating science that relates to almost anything that someone does.

Statisticians have been described as number crunchers, numerologists, and numbers people. Merriam-Webster states: "The statistician is an expert in the preparation and analysis of statistics." From the ASA, a statistician is a person who applies statistical thinking and methods to a wide variety of scientific, social, and business endeavors. The ASA definition is a little more specific than other definitions, but the term "statistical thinking" does raise questions that have been addressed over the years.

None of the definitions or perceptions probably matter if statisticians are treated as scientists working with other scientists.

MW: How have the roles of the statistician and the statistician's circle of collaboration changed in the last 10–20 years? What remains the same?

SS: I have certainly experienced changes over time, but they may be more situational than temporal. When I started out as a statistician, in 1984, I was part of a team at the North Carolina State University (NCSU) Agricultural Research Service. I interacted with the scientists daily and had a lot of influence regarding experimental design and data collection. I was often asked to co-author publications and invited to attend and speak at conferences and workshops. It was the ultimate in collaboration.

I left the university environment in the early 1990s and took a position as a senior statistician with a clinical research organization (CRO). This move changed the way I collaborated. The pharmaceutical companies had the statistical "brains" and needed the CROs as their "worker bees," so being able to follow others' leads was key to success as a statistician in the early CRO business.

With the boom of small to midsize biotech companies in the mid- to- late 1990s, the focus of CRO statisticians began to shift from being the extra personnel in large pharmaceutical companies to becoming the statistical experts for small pharmaceutical companies. When I transitioned to being the sole statistician in a small biotech company, my responsibilities shifted to strategic planning and product development. I was able to contribute my statistical expertise and to learn more about the other components of drug development, such as regulatory processes and manufacturing.

Today, [I am] a consultant, [and] my clients expect me to give them guidance rather than hands-on analyses.

ABOUT THE PANELISTS



Daniel H. Mowrey served as a statistician for 36 years in the animal and human health pharmaceutical business at Eli Lilly and Company. He has worked in many scientific areas and collaborated with chemists, toxicologists, biologists, microbiologists, veterinarians, medical doctors, meat scientists, nutritionists, engineers, regulatory, safety, and quality professionals.



Jonathan Potts has spent the last 10+ years in the credit risk industry, from creating custom predictive models for the telecommunications clients at Equifax to developing credit authorization strategies for a major home improvement retailer at Citigroup. Jonathan recently transitioned to Experian where he is a Senior Data Scientist in a the newly formed Analytics on Demand vertical.



Susan Spruill is a consulting statistician with 30+ years of experience in clinical trials and regulatory requirements. An accredited professional statistician (PStat®), she also serves on the Editorial Board of *Significance* magazine, published by Wiley-Blackwell on behalf of the Royal Statistical Society and the American Statistical Association.



Walter Stroup is a professor in the Department of Statistics at the University of Nebraska, Lincoln, where he served as the chair from 2001–2010. An ASA Fellow, he teaches courses in design and analysis of research studies and in statistical modeling to graduate students in statistics and in a variety of disciplines that use statistics, and has collaborated with researchers in agriculture, various science disciplines, and education.



Michiko I. Wolcott is the managing partner and principal consultant of Msight Analytics, a management consulting firm specializing in data and analytics. Backed by 20 years of experience in leading analytical project execution and delivery worldwide, she helps organizations of all types and sizes in the development and implementation of enterprise capability and effectiveness in data and analytics.

MW: What are the key barriers to successful collaboration between statisticians and others? How does the current culture within statistics [affect] our ability to collaborate, and how does it need to change?

SS: As statistical collaborators, we should not assume every prospective collaborator wants the same thing. Phrases like "data are data" may be true, on the surface, but how you explain the results to your collaborators makes all the difference. While the statistician may be the statistical expert, the collaborator is a scientific expert. We should always be willing to work as a team to solve a problem and learn from others.

DM: Many statisticians today want collaborators to change how they collaborate with us. However, collaborators probably don't know how to change because they only know the way they were "brought up." Maybe it is up to the statistician to make the first move by borrowing from *The Seven Habits of Highly Effective People* by [Stephen] Covey.

Let's begin by being more proactive—by having statisticians convey the same message when it comes to studies, so collaborators don't get mixed messages on what is and what is not acceptable.

Let's stress to journals that, although significance might be important in some cases, it is what is learned from following a well-designed scientific process that is more important. Let's stress the importance of beginning with the end in mind so there is no need for guessing what to look for when the study finishes up.

Let's stress the need to put first things first, such as ethics, excellence, and quality.

Let's think win-win with our collaborators, so they can see that we care about their work. Let's seek to understand the problem, listen more, and speak less. Let's work together to improve and be willing to adjust if things need to change.

MW: What are some reasons for the singular focus on analysis execution today? How has the democratization of analytics [affected] this, and how do we bring the focus back to the scientific process?

JP: The return on investment is a key reason. I think [people once saw] analytics as an add-on to traditional services, especially in the credit analytics world. Now that there has been a substantial investment in Big Data and analytics, many businesses are looking for a monetary return in each analysis. This can be troublesome, since not all analysis leads to a solution.

DM: To most people, statistics or statistician still have the same meaning, regardless of data science or predictive analytics. The scientific community still believes that statisticians work with data, so what has changed? Not much, except the data sets are larger. However, within the statistics community, and among statisticians, there is a greater awareness about what is being done in the name of statistics and who is being classified as a statistician.

Today, we have data scientists, predictive analysts, and machine learning analysts, and others who some classify as statisticians while others would say [they are not]. It is something statistics has been struggling with because it is associated with change, and not many people like to deal with change.

JP: Often, statisticians are brought on board a project after data have been collected and projections have been promised. One way to bring back the scientific process is to analyze projects that use statisticians in the pre-planning/design phase of a project versus those that did not. Hopefully, the added value can be seen immediately in the projects where statisticians were included at the beginning. Having the statisticians participate in planning will make them aware of the objectives and hence be able to propose the proper analyses.

DM: When others seek the collaboration of a statistician, they generally have a problem and want statisticians to come to their rescue, especially if there was not any forethought in the study design or the analysis to be conducted. The real question is: How do we change this paradigm so their learning is a more-robust scientific process that involves statisticians at the beginning of the thought process?

MW: What do statisticians bring to the table? What do our collaborators bring to the table?

SS: We are problem solvers. The collaborators bring us interesting problems, not tasks. Even apparent tasks, such as reviewing a manuscript, can become a problem to solve if we can see a better experimental design or way of rewording a statement. We must always be on the lookout for how we can make the outcome better or more meaningful for the collaborator, even if the outcome may not be what they originally hoped.

JP: We all must recognize that everyone brings a different important skill set to the table. We must be willing to accept input from various angles and be flexible. Many times, collaborators (e.g., sales consultants) may have business insight into a problem that can explain possible trends in data that the statistician may question.

MW: What are the gaps not taught in schools today? How do we better prepare students?

WS: First, integration of scientific thinking (or empirical decision making) and statistical reasoning. Courses on statistical methods focus on the mechanics of methods and their technical interpretation; the hypothesis test and associated p-value is an example, and is the one that originally motivated these sessions.

However, these courses make little attempt to ask what the scientist or decision-maker really needs to answer the question or to make a decision. Introductory courses have to do a much better job of integrating these two.

Second, analysis is overemphasized compared to planning. Walt Federer once observed that most design of experiments courses are really ANOVA courses with little, if any, of the planning aspect of

design. Most experienced statistical and scientific collaborators agree that statistics can add important—often crucial—contributions before any data are collected. Too often, our courses fail to reflect this. Courses that do preach "you must plan" frequently fail to teach students how to plan.

Third, non-statistics majors and statistics majors are not taught how to be effective collaborators. Collaboration has to be taught systematically at universities. Non-majors are not taught how to work effectively with statistical partners. Even when statistics majors are taught to consult, they are not taught key aspects of how to be an effective partner in a scientific investigation.

The curriculum has to be revamped and provide opportunities for collaboration. It has to cover how to "think" like other scientists, or to put another way: We need to teach statistics majors how scientists think.

In addition, statistics majors should be shown how to think of a statistical analysis as a means to an end (answering a scientific question) and not as an end in itself. Consulting classes tend to emphasize the human side of consulting, and statistical methods courses tend to emphasize the mechanics and supporting theory of data analysis. Both are important and should be considered essential parts of the curriculum, but a missing part is how to get inside a scientist's or decision-maker's head and understand what they need. Learning how to do this—integrating what they need with what statistics can offer—is all too often a missing third element.

We need to rethink the curriculum of statistical methods courses to include how to plan and how to collaborate. People in our discipline have been doing a lot of thinking about the undergraduate introductory statistics course, and one of the distinctions they make is between formulas that teach statistical thinking and formulas that merely show how to hand-calculate what people in the real world will use a computer to do. The former has learning value. The latter has no place in a statistical methods course.

Collaboration is something that you learn by doing, not by watching and listening. There are several ways one can approach collaboration. At the University of Nebraska, we partner majors and non-majors in our graduate-level design and analysis course. We have them do a project in which the non-major proposes an experiment. The major and non-major students plan the experiment and do a precision and power analysis in such a way that we can generate simulated data according to the design and anticipated mean and variance parameters the project team gives us. Then they analyze the data and do an oral and a written report.

This way, they get the experience of doing a project from beginning to end.

Another idea is to pair statistics graduate students with graduate students in other disciplines doing their thesis or dissertation research, [with] both students under the supervision of their respective advisors (e.g., the statistics professor could be the outside member on the non-major's committee and vice versa).

Whatever is pursued, always remember that collaboration is doing.

Faculty in statistics departments need to appreciate that planning and collaboration are critical components in research. As the line goes from a Pogo cartoon some of us are old enough to remember, "We have met the enemy and he is us." Many statistics faculty members do not fully appreciate the importance of integrating scientific and statistical thinking. Many mathematically trained faculty members are not comfortable teaching statistics as opposed to mathematical statistics.

George Box once said, "We can choose to be second-rate mathematicians or first-rate scientists." I may be preaching to the choir, but our statisticians teaching methods courses have to reflect what Box obviously believed to be the right choice. All too often, they do not.

MW: To plan and collaborate, how do we better hire, train, and retain faculty?

WS: University statistics departments are all over the map in terms of whether collaboration with scientists in allied disciplines is valued, supported, and rewarded.

In land grant universities, there should be no question that collaborative research is a core mission of the statistics department, and faculty from statistics and collaborating disciplines should be encouraged and rewarded for team efforts that result in good science. All too often, however, this does not happen.

We live in an era where single-author publications, big grant dollars as lead principal investigator, and high-profile research solely by statisticians and solely about statistical topics are the currency of the realm. There is a need for this, but if this is all that counts, statistics risks becoming insular, at least in university research, and good science loses.

In many universities, collaboration is increasingly seen as something for non-tenured faculty ("professors of practice" is the University of Nebraska term) and sometimes consulting is not even considered a legitimate faculty level activity.

Often, new assistant professors get the message that collaborative research is professional suicide—if you want tenure, avoid collaboration. To some extent, this is a prejudice many of our colleagues who sit on promotion and tenure (P&T) committees share and perpetuate. If we allow applied statistics to be synonymous with second class, we will have a self-imposed problem. Faculty cannot teach how to plan and collaborate unless they have experience and are actively engaged in at least some interdisciplinary collaboration.

I have spent my entire career at land grant universities, so I cannot speak about what goes on at other universities with the same authority. I suspect that in many cases, courses on statistical methods are taught by instructors whose primary training is in mathematics, and who may in fact be mathematics faculty. I do not think we should look down at them, but I do think that the ASA should be proactive in providing instructors who fit this description with some form of in-service training to ensure that they teach statistical methods and scientific *reasoning* courses rather than the rote application of statistical techniques.

Finally, some years ago, the ASA put a lot of effort into guidelines for introductory statistics courses (GAISE). Each year at JSM, there are dozens of sessions on teaching introductory statistics courses to high school students and first-year college students. We need to put similar effort and energy into statistical methods curriculum and courses for graduate-level scientists in training.

DM: Hiring, training, and retaining faculty is a good start to addressing some statistical deficiencies in collaboration and planning. However, what about maintaining our organizations' knowledge?

It has been said that an organization's knowledge is inherently bound to the organization's culture. For many years, our culture has been following sound scientific principles and trying to promote those principles to the people with whom we interact. In years past, we have had champions who promoted these scientific principles and kept them in front of our organization—G.W. Snedecor, W.G. Cochran, G.E.P. Box, J.W. Tukey, and D. F. Cox, to name a few. I am not sure those champions still exist, but we need these champions now more than ever; without them, our organization(s) will falter and eventually fail.

As we look around, do we see our statistical organizations—universities in particular—as still maintaining the knowledge that they had in the past; a knowledge that had a strong foundation in statistical principles and guided people in being scientifically sound scientists? Will someday that knowledge become a thing of the past, maybe forgotten forever? Let's make sure that this does not happen.

MW: What can we do to improve planning and collaboration as individual statisticians? As a profession?

SS: We need to listen and ask questions. The statistician must be willing to read/learn more about the collaborator's field of science, so questions and solutions can be made relevant. Collaborators desire someone who is strategic, reliable, and well-spoken.

We should also be proactive and understand that we are all in this together. We may want our collaborators to change how they work with us, but they may not know what that entails.

WS: We need to be active in getting our statistical methods courses right. We also need to be active in communicating with faculty in disciplines that are consumers of statistical methods what the problems are, and why and how we are revamping statistical methods courses to address them.

We need to be active in non-statistical professional societies. We should work with their leadership to agree on statistical best practices for research and publication in their journals. We should seek to understand, appreciate, and address their needs and we should be willing to provide continuing education.

We need to identify faculty members who "get it" and use their expertise to communicate to departments on campus who are consumers of statistical methods. These faculty members could present the latest developments in statistical practice, do seminars, and attend new faculty orientations.

We need P&T committee leaders who make sure the university culture is friendly to statistics faculty who collaborate. If none of the tenured university faculty are so engaged, then the university's ability to educate and train the next generation of statistical collaborators is imperiled.

Finally, many researchers say, "I would love to have a collaborating statistician on my research team. Where do I find one?" This is a real problem at colleges and universities that do not have a culture of collaboration, do not reward statistics faculty for collaboration, and/or do not have a viable consulting/collaboration infrastructure. Could the ASA develop some viable mechanism to connect researchers caught in this quandary with statistical help?

Mobile Apps versus Web Browsers: A Comparison of Self-administered Survey Platforms

Briana Cameron, 23andMe Research Team, and Robert Gentleman

The research program at 23andMe, Inc., a personal genetics company, studies a variety of health and medical conditions using customers' genetic and self-reported health information. While genetic information is captured through the company's genotyping service, surveys remain one important means of collecting phenotypic information.

Customers who consent to participate in 23andMe's research program are given the opportunity to answer questions about a wide variety of topics, including sleep, eating behaviors, and cancer. These self-administered surveys create a rich database of phenotypic information that, when combined with a customer's genetic information, forms the basis of the research program at 23andMe. At this time, 23andMe has more than 4 million genotyped customers and more than 1 billion phenotypic data points, for one of the largest data sources on human health.

For more than 10 years, 23andMe has used a custom-designed survey methodology platform that delivers survey content and collect customer responses via web browsers. As smartphones have become ubiquitous, researchers wanted to explore the opportunities and pitfalls of data collection based on a similar infrastructure that is presented through an app on customers' phones.

Collecting survey data via mobile phones has important considerations, however. There may be a different customer demographic that engages with surveys through the app versus through the web. Furthermore, while the accessibility of mobile phones may create an easy opportunity for customers to participate in research, the usage of phones in potentially distracting environments, such as while commuting or waiting for a friend to arrive for dinner, may ultimately affect survey completion and data quality.

As we collected data across the different platforms, we wanted to ensure that survey responses did not depend on the platform used to complete the survey and that it was appropriate to combine survey results across the different groups. Understanding the possible differences between these data collection platforms would help inform appropriate analysis and interpretation of 23andMe survey data.

Methods

In an attempt to understand differences between web-based and mobile-based data collection, we decided to investigate some of the responses to a series of four surveys administered on both the 23andMe website and 23andMe mobile app. The surveys cover a wide variety of topics, including socioeconomic status (SES), tobacco use, allergies, and caffeine intake, and were chosen because they were available to all customers on both platforms.

We considered new customers who registered with 23andMe between January 1, 2016, and November 15, 2017, for analysis. The analysis population consisted of customers who had consented to participate in 23andMe research and had taken at least one of the four surveys of interest within those dates. Of these customers, 461,482 took at least one of the four surveys selected for analysis and 292,626 customers responded to more than one of the four studied surveys. A summary of the length of each survey, including number of questions asked of every customer (i.e., top-level questions), number of total possible questions, median number of questions answered, and median time to complete survey, is included in the Supplemental Material at *http://chance.amstat.org/2018/11/supp-material/*.

While our main interest was in the difference between responding to surveys through an app versus through a web browser, we also investigated potential differences within the population of app-users. Android phones and iOS phones use slightly different apps

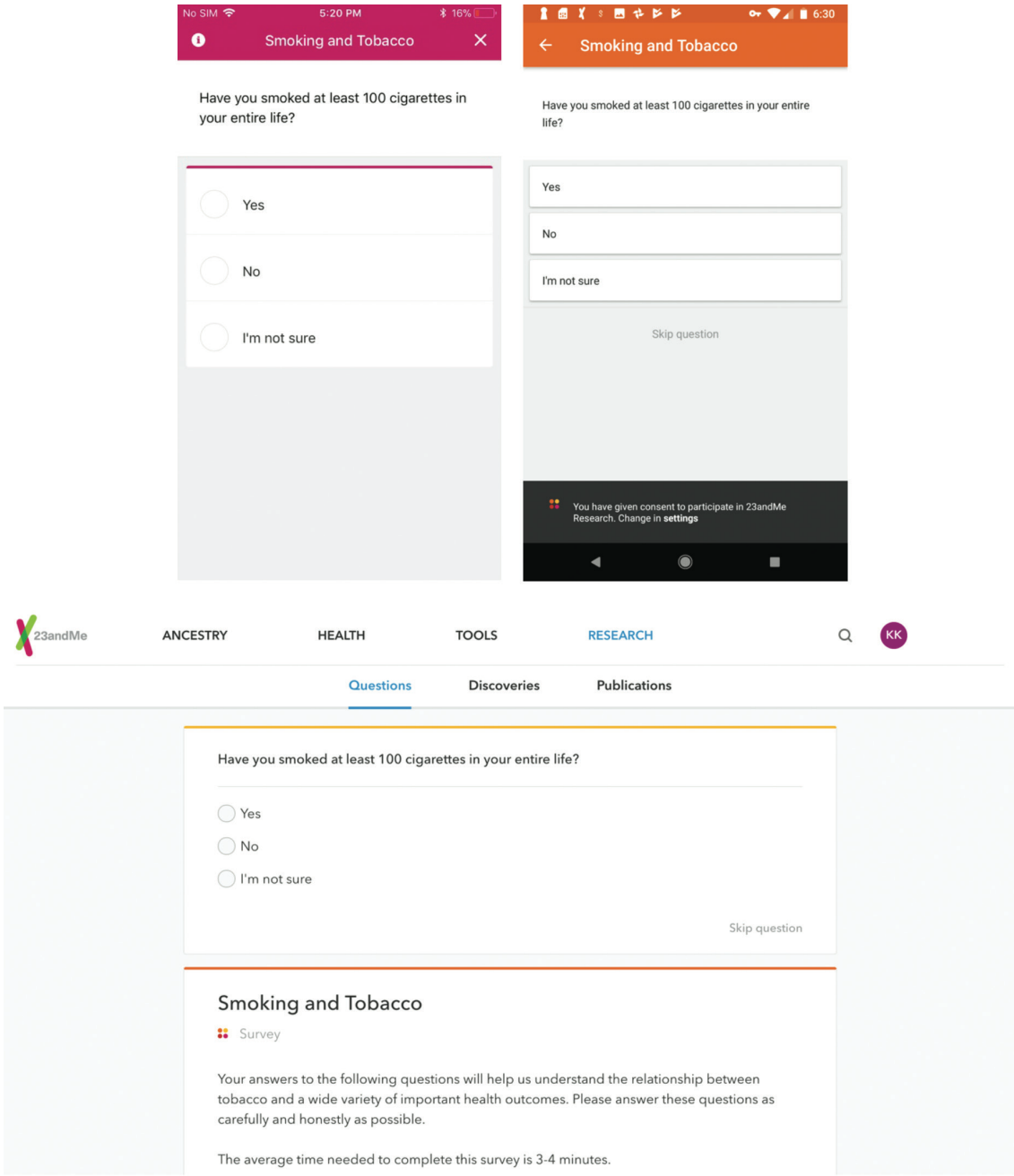


Figure 1. Screenshots of a question from the tobacco survey on both mobile apps (top left: iOS app, top right: Android app) and desktop website.

(Figure 1), and we were interested in determining whether there were differences between these two groups.

There may also be important subgroups among customers answering surveys on the web. Customers are able to access these web surveys through either a desktop browser or a mobile web browser, which have important differences in terms of screen size and the environment in which they are accessed. The information regarding a customer's browser was not easily accessible, however, so we were unable to make the distinction between desktop and mobile web in this analysis.

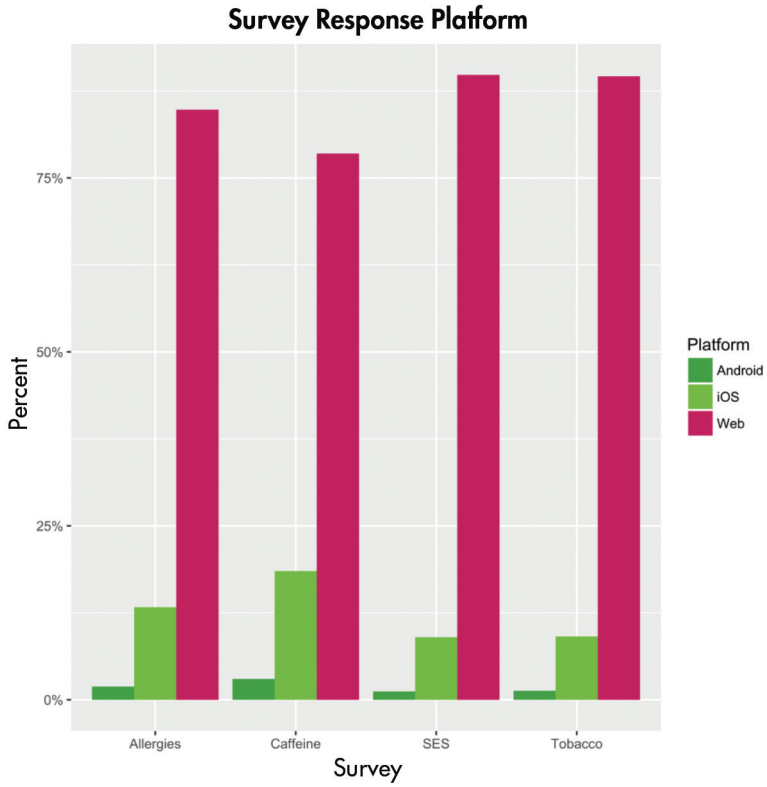


Figure 2. Percent of customers using each survey platform.

Questions of Interest

The first question of interest was whether the population being reached via mobile surveys was similar to the group of customers who typically participate in research through the 23andMe website. To investigate this question, we considered the key demographics of age, sex, and race. We also investigated possible socioeconomic differences between these groups.

In addition to identifying possible demographic differences, we were also interested in identifying variation in survey-taking behaviors between the two cohorts. In particular, we wished to test the hypothesis that customers engaging in the survey through the mobile app may be more distracted than those answering questions through the website.

One measure of distraction may be survey drop-off. Of the customers who initiated the survey, how many actually complete it? Increased distraction may also be measured by an increased likelihood of skipping questions. A variety of question types were selected for analysis and the likelihood of skipping these questions was assessed among both mobile and web survey-takers as an additional measure of distraction.

Results

Across all four surveys, the majority of customers opted to take the survey on the web platform, with only 10.2–21.5%* of customers responding via the mobile app. Of the customers responding via the mobile app, the majority used the iOS App, while only 11.8–14.1% of mobile app customers using the Android app in each survey (Figure 2). Customers also tended to stick to their preferred method of answering surveys; of the customers who did complete more than one of the four studied surveys, 95.8% completed all surveys on the same platform.

Demographics

Do survey data collected via the mobile app come from a different demographic from survey data collected via the website? To answer this question, we considered the key demographic variables of age, sex, and race.

Customer age is one domain in which differences might be expected between mobile app and web survey respondents. To investigate this, we used the median age of the research database to split customers into a younger group

*"Reported range" refers to the range of observed percentages across the four surveys included for analysis. This is true of all reported percentage ranges.

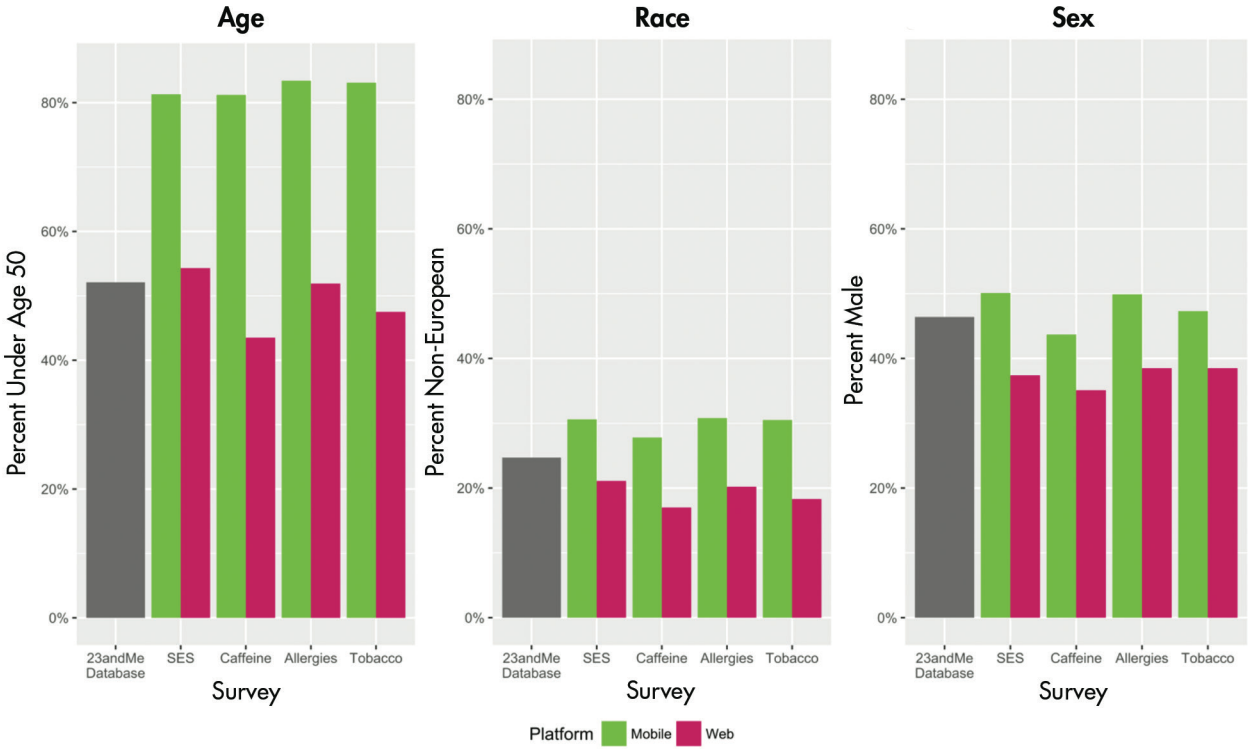


Figure 3. Race, sex, and age breakdown of respondents in each survey platform group; gray bar indicates the demographic of the entire 23andMe research database.

(customers under 50 years old) and an older group (customers aged 50 or greater). For each of the studied surveys, we tabulated the proportion of younger customers and older customers. In all surveys, mobile app respondents were significantly more likely to be younger, with 81.2–83.4% of customers falling below the 50-year-old mark. In contrast, only 43.5–54.3% of web respondents fell into the younger age group (Figure 3). No significant differences in age were observed between the iOS and Android app groups.

We also noted differences in terms of the race of respondents. Customers engaging in the survey through the mobile app were more likely to be non-European than their website counterparts. While 78.9–83.0% of web respondents were of European ancestry, customers of European ancestry accounted for only 69.4–72.2% of mobile app respondents (Figure 3). In general, however, the majority of survey responses across both platforms came from customers of European ancestry, consistent with the distribution of ancestry in the 23andMe database as a whole.

Similarly to age, we did not observe any differences in race between Android and iOS app users.

Finally, we noted sex differences between the two platforms. Across all platforms, respondents were more likely to be female, although this trend is exaggerated in the group of customers responding via the website (61.5–64.9% female) (Figure 3). A more-even distribution between the sexes was observed in the mobile app group, where 49.9–56.3% of respondents were female. Interestingly, we also found differences in the sex of respondents within the group of customers who completed the survey on the mobile app. Customers who used the Android app to complete the survey were more likely to be male than those completing the survey via the iOS app, although both iOS and Android app users were more likely to be male than web respondents.

Taken together, customers who respond to 23andMe surveys on the mobile app are more likely to be younger, non-European, and male than customers who take surveys on the website. This observation may help researchers collect data

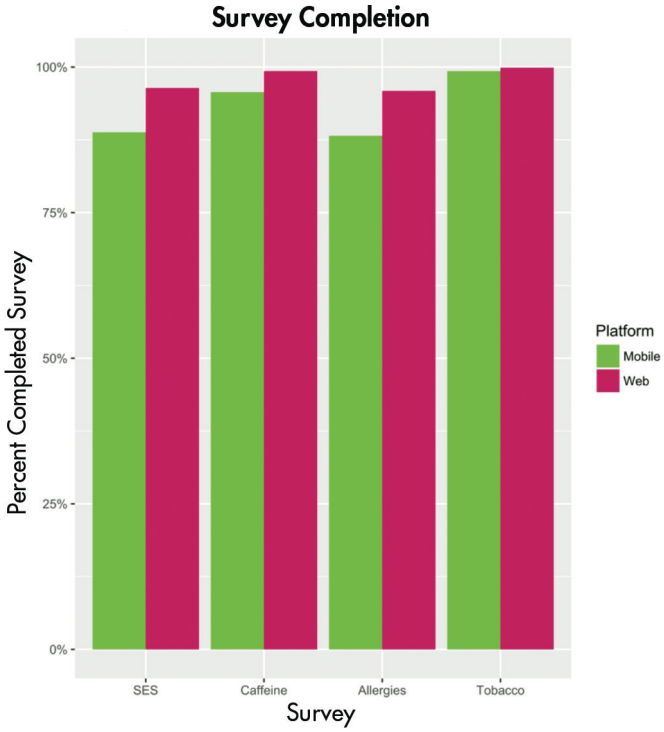


Figure 4. Percent of customers completing each survey by survey platform.

from groups that may not be well-represented in website-based data collection, including men and non-European respondents.

We were also interested in potential socioeconomic differences between web-based survey responders and app-based responders, after adjusting for the demographic differences noted above. To do this, we looked at a series of questions from the socioeconomic (SES) survey related to income, education, home ownership, residence type, relationship status, and having worked at least 50 weeks in the past year.

After adjusting for age and sex, mobile app users were significantly more likely to report having worked at least 50 weeks of the last year, more likely to be in a higher income group (annual income > $75,000), and slightly less likely to own their homes (either with or without mortgage).

Survey-taking Behavior

The fact that mobile surveys tend to reach a demographic distinct from that captured through traditional web-based surveys has important implications for researchers trying to collect data from a diverse and representative population. It is also important to understand how data collected via these two media may differ, however.

Surveys administered on mobile phones may be more easily accessible in certain demographics, but may also be taken in environments with greater distractions for customers. The context in which a survey is taken may result in important differences in survey engagement between the two platforms.

Completion data for the four surveys of interest seem to support this notion. While the vast majority of customers who started the survey did ultimately complete it (94.7–99.8%), there were differences in completion rates between the two survey platforms. Of the customers initiating any of the four surveys, 96.4–99.9% of those taking the survey on the website ultimately completed the survey, while mobile app respondents completed the survey 88.2–99.3% of the time (Figure 4). For both survey platforms, completion rates were also closely tied to the length of the survey, with longer surveys experiencing a greater rate of mid-survey drop-off.

Another way in which decreased survey engagement may manifest is the likelihood of skipping questions. For all survey items, customers have the option to

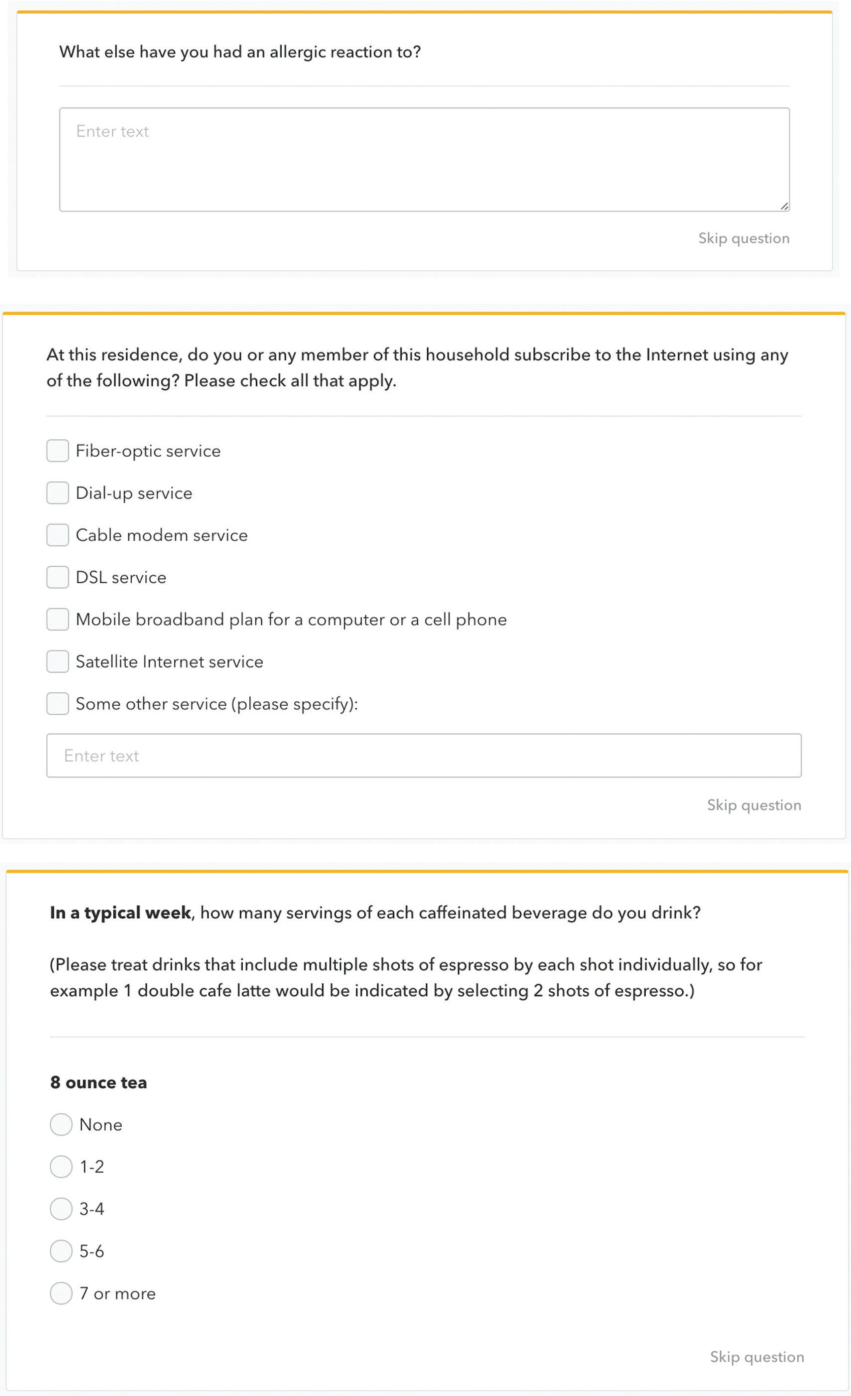


Figure 5. Screen shots of the different question types: free text (top), checkbox (middle), radio buttons (bottom).

abstain from providing an answer and can, instead, choose to skip the question. One concern may be that the distractions in mobile-based surveys will make mobile app users more likely to omit answers to certain types of questions.

To investigate this issue, we selected a set of questions and determined the proportion of customers opting to skip each question within each survey platform. It is noteworthy that these surveys included a variety of question types, which place different demands on the customer (Figure 5). Some free-text questions require customers to input text as the response (i.e., "What else have you had an allergic reaction to?"). Other checkbox questions

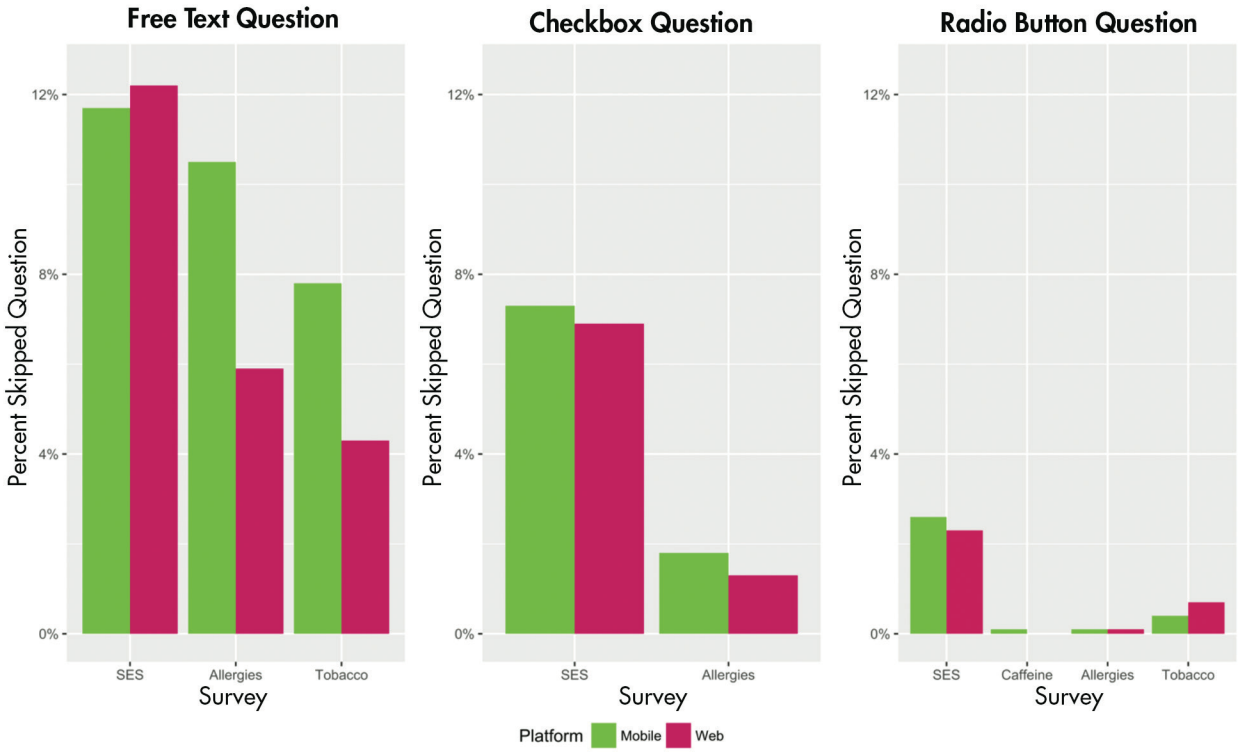


Figure 6. Percent of customers skipping each question type broken down by platform.

ask customers to select all applicable answer options from a set list (i.e., the types of Internet services that a household subscribes to). Radio button questions may ask users to select a single option from a set of defined answer choices (i.e., "In a typical week, how often do you drink tea?").

Considering that each of these questions types places different demands on survey participants, we decided to look at the probability of skipping separately for each question type.

The questions with the highest customer demand are those with free text responses. Unsurprisingly, customers using both survey platforms were more likely to skip these types of questions than the other question types, with 4.3–12.2% of customers declining to provide an answer (Figure 6). A comparison of skipping patterns between web users and mobile app users indicate that there may also be differences in likelihood of skipping between these groups. In 2 out of 3 free text questions analyzed, mobile app users were almost twice as likely to decline providing an answer than web users (10.5% vs. 5.9% skipped, 7.8% vs. 4.3% skipped) (Figure 6). This observed difference may be evidence of distraction among mobile-app survey-takers or may be due to the inconvenience of typing on mobile phones.

We did not see any major difference in skip rate for the last free text question analyzed (11.7% skipped on mobile vs. 12.2% skipped on web). It is worth noting that this last question came from the SES survey, which generally had much higher skip rates among web users than the other three surveys included for analysis, potentially due to the sensitive nature of the questions asked in this survey.

The second major question type considered was the checkbox version, which asked customers to select all applicable answer choices from a set list. Overall, fewer people opted to skip this type of question than the free text questions (1.3–7.3% skipped) (Figure 6). This question type was not often used in the surveys included for analysis; of the two checkbox questions analyzed, we did not observe any difference in skip patterns between mobile and web-based responses. We also considered the number of boxes selected for each of these checkbox questions; again, we observed no difference between web and mobile app respondents in this regard.

The final, and most popular, question type is the radio button,

in which users select a single option from a specified list of answer choices. Of the three question types considered, radio button questions are the simplest for users to answer. This is evidenced by an extremely small skip rate: Only 0.1–2.6% of customers tend to omit answers to these questions (Figure 5). The low skip rates were observed across both platforms, and we did not find any meaningful differences in skip rates between web respondents and mobile app respondents.

Conclusions

The use of mobile applications to capture research data efficiently is a promising technique for researchers. Mobile phones are carried by most people throughout the day, creating an easily accessible avenue for customers to participate in research. The use of mobile phones also allows for collecting more context-dependent data, such as location and weather, that might be of interest to researchers.

Our goal was to understand whether important differences existed between data collected through the mobile app and data collected through the more traditional web surveys. Our analysis demonstrates important differences in demographics between customers engaging in research on each platform. In particular, the use of mobile apps may allow more data to be collected from younger, male, and non-European respondents. The ability to access information from this distinct demographic group makes the use of mobile surveys an important complement to the traditional web-based survey platform.

While surveys administered via mobile apps may increase the data collected from certain populations, our analysis also illustrates important caveats to this approach. The convenience that makes these surveys accessible to a broader demographic may also contribute to decreased engagement in the research experience. In particular, we observed that customers using the mobile app are more likely to end the survey prematurely and to skip answering free text questions.

Based on these observations, we may ultimately collect fewer data per person via the app versus via the website. However, with the growth of mobile-app–based data collection, the use of mobile apps as a research tool may continue to be an important complement to many of the more-traditional web-based data collection techniques.

Acknowledgment

The authors thank the research participants and employees of 23andMe, Inc. for making this work possible.

About the Authors

Briana Cameron, a scientist working in biostatistics, joined 23andMe in 2017 and is responsible for the development and implementation of statistical methods to analyze 23andMe's complex phenotypic data set. Before joining 23andMe, she received her PhD in biostatistics from Yale University, where her research focused on the development of novel clinical trial methodology and the evaluation of the role of patient choice in evaluating treatment efficacy.

Robert Gentleman joined 23andMe in 2015 as vice president of computational biology. In this role, he focuses on exploring how human genetic and trait data in the 23andMe database can be used to identify new therapies for disease. He has received the Benjamin Franklin Award and recognition for Open Access in the Life Sciences from the Bioinformatics Organization, and is a Fellow of the International Society for Computational Biology (ISCB). Gentleman is one of the originators of the R programming language, along with Ross Ihaka at the University of Auckland, and was one of the founders of the Bioconductor Project.

The Lunacy of Motorcycle Mortality

Donald A. Redelmeier and Eldar Shafir

Motorcycle riding is a popular way to exchange longevity for excitement and now accounts for more deaths in the United States than traditional diseases such as asthma, appendicitis, or meningitis.[1] A one-hour ride on a motorcycle has a greater overall risk of death than a lifetime of commercial airline flights.

Unlike the declining trend in mortality for car drivers, the annual number of motorcycle fatalities has more than doubled over the past 20 years.[2,3] Motorcycle crashes now total nearly 5,000 deaths annually, accounting for $6–$12 billion in societal costs and 1 in 7 road fatalities.[4] New public health policies partially explain the increase in motorcycle deaths, such as the repeal of universal helmet laws in 27 states.[5] Motorcycle crashes are often blamed on simply bad luck.[6]

The full moon, for its part, has attracted attention since antiquity and generated a plethora of superstitions about adverse events that include deviant behavior, violent crime, medical illnesses, nighttime insomnia, altered fertility, and supernatural misfortunes (i.e., lycanthropy).[7,8]

A century of research, however, has failed to validate such claims. Indeed, most studies claiming paranormal effects contain basic design flaws, isolated studies have been refuted directly, and magical beliefs have been shown typically to reflect statistical fallacies such as confirmation bias.[9,10,11,12,13]

Superstitions about the adverse effects of a full moon are now a target for public ridicule.[14] Not surprisingly, no prior study has tested the effect of a full moon on motorcycle fatalities.

We hypothesized that motorcycle fatalities might increase at night under a full moon because human attention is limited, naturally drawn to attention-grabbing events, and a factor in traffic crashes. Glancing at a full moon may take the rider's gaze off the road, for example, and thereby can create a moment of dangerous inattention.[15,16] The light from a full moon also illuminates the entire landscape and thereby creates potentially distracting effects from extraneous incidentals in the surroundings.[17]

---

[1]National Vital Statistics Reports. 2012. Deaths: Preliminary data for 2011. Washington, D.C.: Centers for Disease Control, [61, 6]. *http://www.cdc.gov/nchs/data/nvsr/nvsr61/nvsr61_06.pdf*.

[2]National Highway Traffic Safety Administration. 2002. Traffic safety facts: motorcycles. Washington, D.C.: U.S. Department of Transportation [DOT HS 812 035]. *http://wwwnrd.nhtsa.dot.gov/Pubs/812035.pdf*.

[3]Hedlund, J. 2015. Motorcyclist traffic fatalities by state: 2013 preliminary data. Washington, D.C.: Governors Highway Safety Association. *https://www.ghsa.org/sites/default/files/2016-11/Final%20Motorcycle%20Spotlight2013.pdf*.

[4]National Highway Traffic Safety Administration. 2002. Costs of injuries resulting from motorcycle crashes: A literature review. Washington: U.S. Department of Transportation [DOT HS 809 242].

[5]Houston, D.J., Richardson, L.E. 2007. Motorcycle safety and the repeal of universal helmet laws. *American Journal of Public Health*, *97*(11):2063–9.

[6]Busko, A., Hubbard, Z., Zakrison, T. 2017. Motorcycle helmet laws and public health. *New England Journal of Medicine*, *376*(13):1208–9.

[7]Sarton, G. Lunar influences on living things. 1939. *Isis*, 30:495–507.

[8]Lilienfeld, S.O., Arkowitz, H. 2009. Lunacy and the full moon. *Scientific American*. Accessed on January 20, 2015. Available at *http://www.scientificamerican.com/article/lunacy-and-the-full-moon/*.

[9]Rotton, J., Kelly, I.W. 1985. Much ado about the full moon: a meta-analysis of lunar-lunacyresearch. *Psychology Bulletin*, *97*(2):286–306.

[10]Eisenburger, P., Schreiber, W., Vergeiner, G., Sterz, F., Holzer, M., Herkner, H., Havel, C., and Laggner, A.N. 2003. Lunar phases are not related to the occurrence of acute myocardial infarction and sudden cardiac death. *Resuscitation*, *56*(2):187–9.

[11]Morton-Pradhan, S., Bay, R.C., Coonrod, D.V. 2005. Birth rate and its correlation with the lunar cycle and specific atmospheric conditions. *American Journal of Obstetrics and Gynecology*, *192*(6):1970–3.

[12]Biermann, T., Estel, D., Sperling, W., Bleich, S., Kornhuber, J., and Reulbach, U. 2005. Influence of lunar phases on suicide: the end of a myth? A population-based study. *Chronobiology International, 22*(6):1137–43.

[13]Otte, W.M., van Diessen, E., Bel,l G.S., Sander, J.W. 2013. Web-search trends shed light on the nature of lunacy: relationship between moon phases and epilepsy information-seeking behavior. *Epilepsy Behavior*, *29*(3):571–3.

[14]Douglas, I. 2010. MPs believe the funniest things. *The Telegraph*. Accessed on January 20, 2015. Available at *http://blogs.telegraph.co.uk/technology/iandouglas/100005822/mps-believe-the-funniestthings/*.

[15]Schweizer, T.A., Kan, K., Hung, Y., Tam, F., Naglie, G., Graham, S.J. 2013. Brain activity during driving with distraction: an immersive fMRI study. *Frontiers in Human Neuroscience*, *7*:1–11.

[16]Klauer, S.G., Guo, F., Simons-Morton, B.G., Ouimet, M.C., Lee, S.E., Dingus, T.A. 2014. Distracted driving and risk of road crashes among novice and experienced drivers. *New England Journal of Medicine*, *370*(1):54–9.

[17]Kahneman, D., Beatty, J., Pollack, I. 1967. Perceptual deficit during a mental task. *Science*, *157*(3785):218–9.

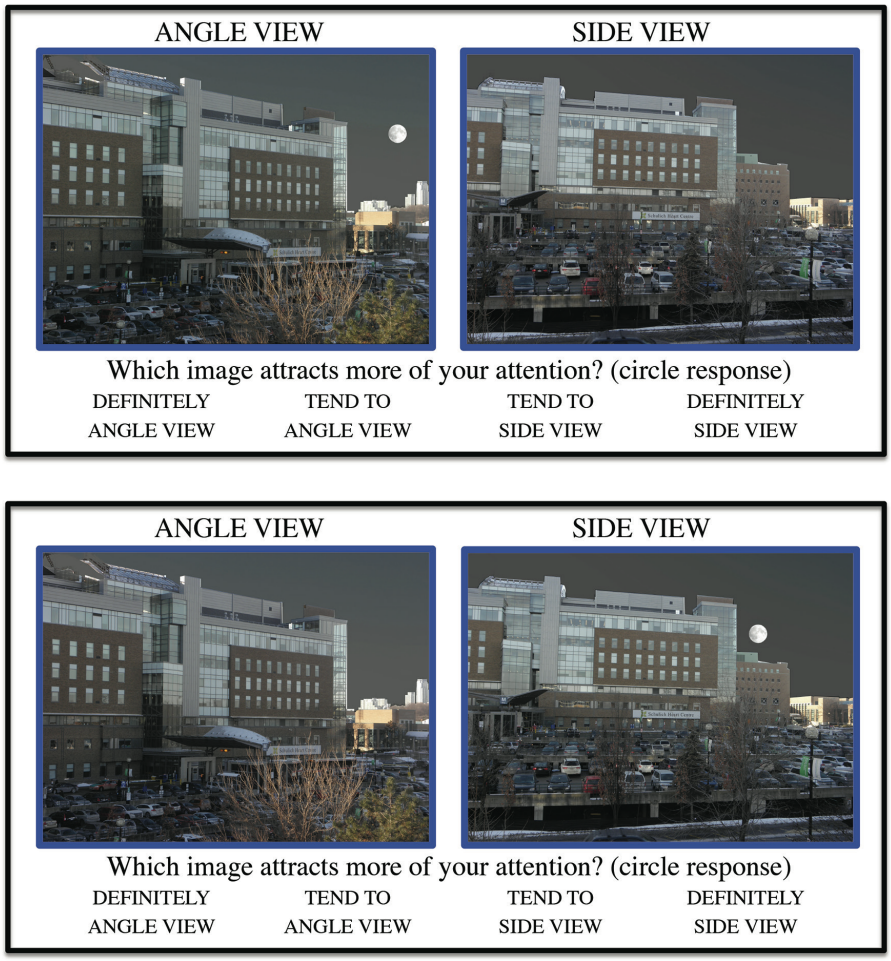


Figure 1. First Test of Distraction from Full Moon.
The two photographs—Angle View and Side View—of Canada's largest trauma hospital are shown in Study 1. The full moon is superimposed on the Angle View and the Side View photos in the top and bottom versions, respectively. The two versions, with photographs sized 3.6 x 4.8 inches, on 8.5 x 11 cardstock paper, were presented randomly to respondents (n = 309) who indicated which photograph attracts more attention. The Angle View photograph was chosen more often (115 of 158) when it contained the full moon and less often (87 of 151) when it lacked the full moon (73% vs. 58%, $p = 0.006$). (Original digital copies of survey materials available on request.)

The full moon, in addition, is seen by all road users and thereby a potential distraction to nearby motorists, pedestrians, and other elements of cross-traffic. A full moon, therefore, might increase motorcycle fatalities without exerting a supernatural force.

Motivating Experiments

To explore this idea, we ran two pilot studies on the full moon's capacity to attract attention. The first study randomly inserted an image of a full moon into one of two photographs of the same building and asked respondents to indicate which image attracted more attention (Figure 1). As expected, the same photograph attracted more attention when a full moon was inserted.

The second study used a psychomotor vigilance test to assess reaction times with random

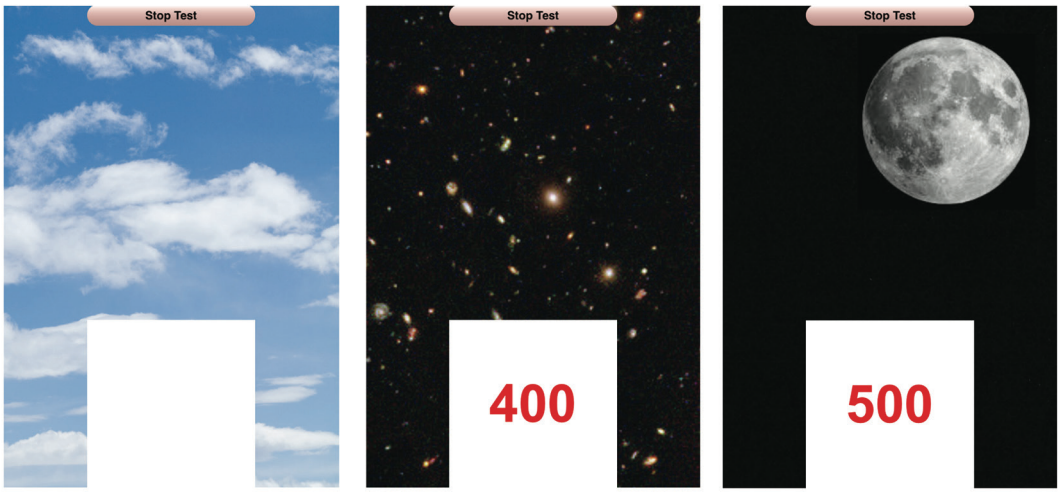


Figure 2. Second Test of Distraction from Full Moon.
Screenshots of Psychomotor Vigilance Test embedded in three background images, with randomized red targets inside white squares indicating elapsed response time in milliseconds (left to right): day sky with clouds at initial resting state with no visible target (not discussed); night sky with stars with active target at 400 milliseconds elapsed time, and night sky with a full moon at 500 milliseconds elapsed time. Targets appeared every 7 seconds on average and background images switched every 30 seconds. Adult volunteers (n = 81) completed the test, which lasted about 120 seconds, on an iPod Touch device. Mean reaction time was slower with a background sky image with a full moon compared to a background sky image with scattered stars (542 vs. 472 milliseconds, $p = 0.002$).

background distractions (Figure 2). As expected, mean reaction time was slower with a full moon background than with a plain night sky background.

Together, these studies motivated our undertaking a larger analysis of real world conditions.

Basic Analysis

Motorcycle crash data for the United States come from the National Highway Traffic Safety Administration, collected from the Fatality Analysis Reporting System, the official database of traffic fatalities in the United States.[18] All 40 years of available data are included with no exclusions (January 1, 1975–December 31, 2014; 494 full moons). The database has been validated in past research, is subject to ongoing scrutiny, remains freely available, and now exceeds 1 million total deaths.[19]

The database has important limitations because it encompasses only fatal crashes on public roads, lacks data on travel distances, and has no measure for individual background exposure.

We defined a "full moon night" as the one time each month when the whole facing surface of the moon was illuminated (as seen from Earth). On rare occasions, a month had two full moons; in such cases, we included both appearances (e.g., June 1 and June 30, 2007). We defined nighttime simply as 1600h to 0800h (16 hours) to include all hours when the full moon might be potentially visible in the sky over the continental United States (including relevant hours of dusk and dawn).

The remaining hours between 0800h and 1600h (8 hours) were defined as daytime and excluded from primary analysis. Fatalities with missing crash times (< 1%) were coded as night (excluding these did not alter our findings).

We defined controls as the same night one week before and one week after the full moon. For example, we matched a full moon on Saturday July 30, 1977,

[18]Fatality Analysis Reporting System Encyclopedia. National Highway Traffic Safety Administration. *http://www-fars.nhtsa.dot.gov/Main/index.aspx. Accessed February 12, 2015.*
[19]National Highway Traffic Safety Administration. 2012. Traffic Safety Facts. *http://www-nrd.nhtsa.dot.gov/Pubs/812032.pdf.* Accessed February 12, 2015.

to control dates of Saturday, July 23, 1977, and Saturday August 6, 1977. Validation of this approach appears elsewhere in previous analyses of traffic deaths.[20,21]

This approach ensured identical time intervals for all comparisons and directly controlled for weekday, seasonal, and yearly trends. It also reduced confounding from unmeasured differences in road design, vehicle conditions, rider training, individual personality, traffic enforcement, healthcare access, and many other influences on crash deaths.[22]

We found the 494 nights with a full moon accounted for 4,494 individuals in fatal crashes, equal to 9.10 deaths per night. The 988 control nights without a full moon accounted for 8,535 individuals in fatal crashes, equal to 8.64 deaths per night. A simple unmatched binomial comparison of the two groups yielded a relative risk of 1.05 associated with the full moon (95% confidence interval 1.02 to 1.09, p = 0.005). The absolute increase totaled 226 additional crash fatalities over the study.

Statistics based on total counts, however, can be misleading because they ignore matching. Instead, a study that is assembled using a matched design should also be analyzed using matched statistics.

Matched Analysis of Matched Studies

Two outlier dates help illustrate the rationale for scrutinizing a matched study carefully. The full moon night July 30, 1977, had 46 deaths and the control nights (July 23, August 6) had 23 and 20 deaths, respectively. Hence, this full moon had double the risk of a fatality. In contrast, the full moon night on January 19, 2011, had two deaths and each of the control nights (January 12, January 26) had four deaths. Hence, this full moon had half the risk of a fatality.

Calculating a simple mean yielded 24 deaths ([46 + 2] / 2) for the full moon nights and 13 deaths ([23 + 20 + 4 + 4] / 4) for the control nights, suggesting a near-doubling of risk associated with a full moon (24 / 13). This calculation, however, is potentially misleading, since July and January are much different in baseline counts.

Hierarchical analyses can preserve the matching of each full moon with corresponding control nights and are readily calculated as conditional odds ratios. For example, the full moon of July 30, 1977, and of January 19, 2011, might be assumed equally important; if so, their estimated relative ratios could be combined by a geometric average to yield an overall relative risk near the null. In general, matched analyses can yield many different conditional odds ratios depending on how much weight is given to individual death counts (e.g., how an absolute increase of 24 deaths compares to an absolute decrease of two deaths).[23] We caution, therefore, that conditional odds ratio estimates of excess risk can be easily calculated and easily misinterpreted.[24]

Traditional approaches tend to consider the person as the unit-of-analysis, assume each death is equally important, and emphasize the imbalance in overall counts. The full moon night of July 30, 1977, therefore, can be summarized as the total count and excess count of motorcycle fatalities relative to controls.

In this case, each of the three nights had a minimum of 20 deaths (denoted as concordant cases) and the full moon night had an excess of 26 deaths (denoted as discordant cases). The excess of 26 deaths can be further distinguished as the 23 deaths unique to the full moon night (denoted as fully discordant cases) and the three deaths shared with one control night but not both (denoted as partially discordant cases). These imbalances can only be appreciated by a matched analysis (Figure 3).

McNemar's test is the classic example of a conditional odds ratio statistic that ignores the concordant cases and considers only the discordant cases in studies with simple 1-to-1 matching. McNemar's test generally yields more-extreme results than a chi-square test that neglects the matching. The data for July 30, 1977, for example, suggest 26 excess deaths on the full moon night and 1.5 excess deaths on control nights ([3 + 0] / 2).

One calculation of a conditional odds ratio, therefore, yields an extreme estimate of 17.3 by focusing only on the ratio of discordant cases (26 / 1.5). More generally, a conditional odds ratio may be

[20]Redelmeier, D.A., Tibshirani, R.J. 2008. Driving fatalities on US presidential election days. *JAMA. 300*(13):1518–20.
[21]Staples, J.A., Redelmeier, D.A. 2018. April 20 Cannabis Celebration and Fatal Traffic Crashes in the United States. *JAMA/Internal Medicine.* doi: *10.1001.*
[22]Tanner, J.C. 1952. Effect of weather on traffic flow. *Nature,* 4290:107.
[23]Martens, E.P., Pestman, W.R., Klungel, O.H. 2007. Letter to the Editor: Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: *Statistical Medicine, 26*(16):3208–10.
[24]Redelmeier, D.A., Tibshirani, R.J. Methods for analyzing matched designs with double controls: excess risk is easily estimated and misinterpreted. *Journal of Clinical Epidemiology* (forthcoming).

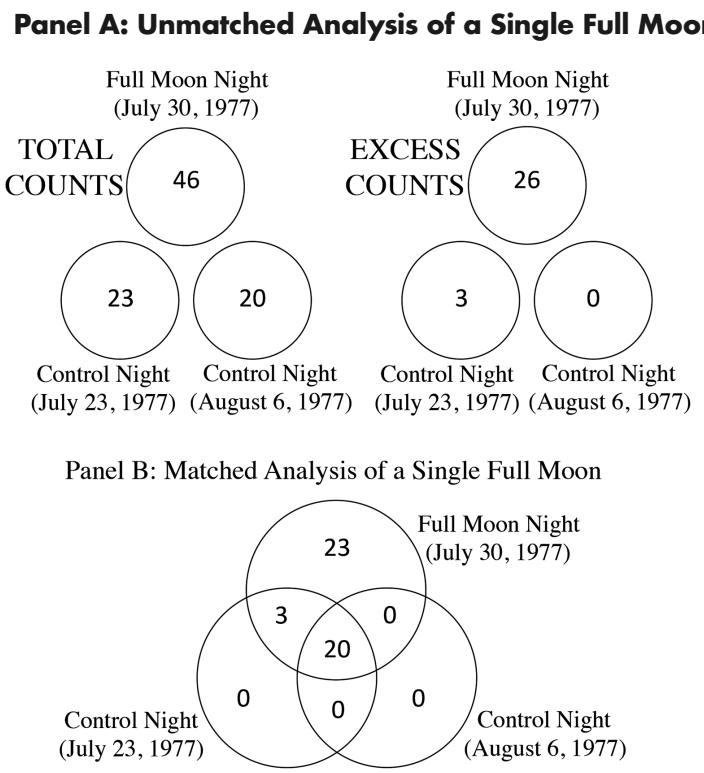


Figure 3. Example Counts for Analysis. *Panel A: Unmatched Analysis of a Single Full Moon.*
Example counts of mortality data from single full moon night (July 30) and corresponding control nights (July 23, August 6). Upper Panel A shows total counts (left three circles) and excess counts (right three circles). Panel B combines total and excess counts together (in Venn diagram), and shows baseline level of concordance and discordance (similar to the "on-diagonal" and "off-diagonal" elements of a 2 x 2 matrix used for McNemar's test). For example, total count on full moon night (n = 46) is sum of concordant cases (n = 20), partially discordant cases (n = 3), and fully discordant cases (n = 23).

useful for testing subtle scientific theory (exploring imbalances), but an analysis of total cases may be more insightful for formulating public policy (addressing totals).

Methods that extend McNemar's test for analyzing 1-to-2 matched studies are now available, as are extensions for outcomes that are accumulating discrete counts.[25] The principle is to distinguish concordant cases (in our case, the number of deaths common across the three nights) from discordant cases (number of deaths imbalanced across the three nights) and further distinguish fully discordant cases (deaths unique to a single night) from partially discordant cases (deaths shared by two but not all nights).

Summary data then yield a conditional odds ratio of 1.26 (95% confidence interval 1.17 to 1.37, $p < 0.001$) associated with the full moon (Figure 4). This conditional odds ratio can be easily misinterpreted; therefore, we based our main analysis on the more-conservative relative risk estimate.

The summary data are another way to contrast unmatched and matched perspectives. An unmatched analysis of total deaths showed 4,494 observed on full moon nights compared to 4,268 expected based on the control nights ([4,323 + 4,212] / 2), a ratio of 1.05 (4,494 / 4,268). Similarly, the excess deaths showed 1,566 observed on full moon nights compared to 1,340 expected based on the control nights ([1,395 + 1,284] / 2), a ratio of 1.17 (1,566 / 1,340). In contrast, the fully discordant deaths amount to 644 observed on full moon nights compared to 540 expected based on the control nights ([553 + 526] / 2), a ratio of 1.19 (644 / 540), and the partially

[25]Redelmeier, DA, Tibshirani, RJ. 2017. A simple method for analyzing matched designs with double controls: An extension of McNemar's test. *Journal of Clinical Epidemiology*, *81*:51–55.

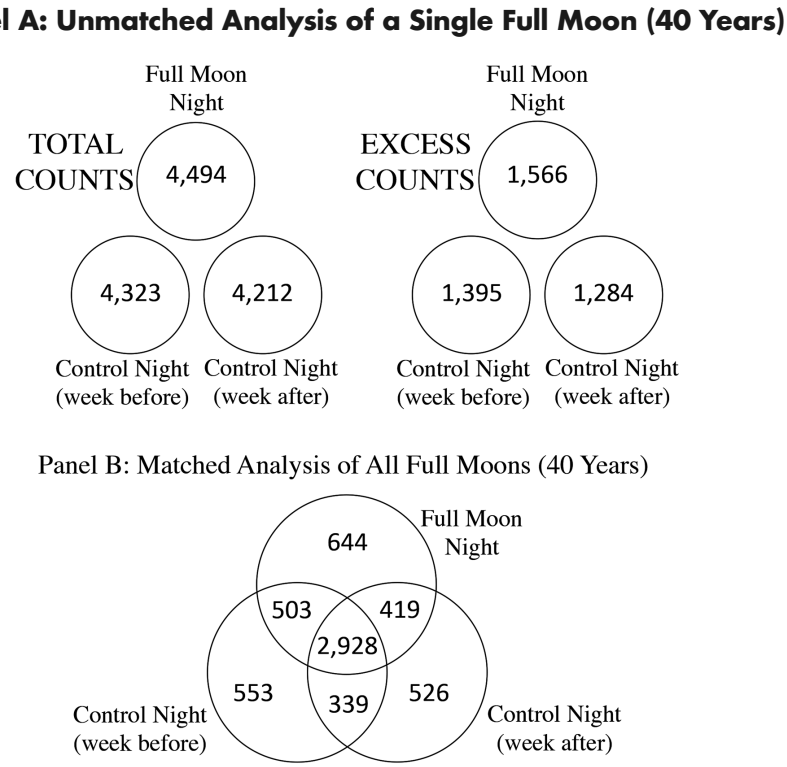


Figure 4: Summary Counts for Analysis. *Panel A: Unmatched Analysis of All Full Moons (40 Years).*
Summary counts of mortality data from all full moon nights and matched control nights over 40 years. Panel A shows total counts (left three circles) and excess counts (right three circles). Panel B combines total counts and excess counts together (in Venn diagram), and shows baseline level of concordance and discordance (similar to the "on-diagonal" and "off-diagonal" elements of 2 x 2 matrix used for McNemar's test). For example, total count over all full moon nights (n = 4,494) is sum of concordant cases (n = 2,928), partially discordant cases (n = 503 + 419), and fully discordant cases (n = 644).

discordant deaths are 461 observed on full moon nights ([503 + 419] / 2), compared to 339 expected based on the control nights, a ratio of 1.36 (461 / 339). Evidentally, the same data yield differing evaluations due to the matching.

Summary Comments

We studied more than 10,000 motorcycle fatalities in the United States over four decades. We found that fatal motorcycle crashes frequently occured at night, particularly under a full moon. The absolute increase exceeded 200 extra deaths attributable to motorcycle riding under a full moon and was five times greater than the total number of deaths from shark attacks in the United States during the same years.[26]

Unlike full moon pseudoscience, the findings were not easily explained as artifacts due to selection bias, missing data, reverse causality, subjective reporting, confirmation bias, known geophysical effects, or random chance.[27, 28, 29] Instead, the data suggested motorcyclists need to ride with more care at night under a full moon.

One obvious limitation of the study is that correlation does not mean causation. In particular, unmeasured factors (e.g., speed, distance, traffic) might contribute to the crash risk associated with a full moon. Our design, however, had no ambiguities in the direction of causality, which means unmeasured factors defined a pathway of risk (mediator) not an alternative source of risk (confounder).

In addition, many lurking variables fluctuate with time and can lead to conservative statistical estimates (e.g., the prevalence of cloud cover was unknown and relative risks could be higher for clear nights). Of course, the moon

[26] *https://en.wikipedia.org/wiki/List_of_fatal,_unprovoked_shark_attacks_in_the_United_States.*
[27] Sackett, D.L. 1979. Bias in analytic research. *Journal of Chronic Disease, 32*(1–2):51–63.
[28] Greenland, S., Morgenstern, H. 2001. Confounding in health research. *Annual Review of Public Health, 22*:189–212.
[29] Hulley, S.B., Cummings, S.R., Browner, W.S., Grady, D.G., Newman, T.B. 2013. *Designing Clinical Research*. Philadelphia, PA: Lippincott Williams & Wilkins.

is also present on intermediate nights, which offers opportunities for future research exploring more-complex Fourier statistics that examine intermediate lunar phases and positions.

Regardless of mechanisms, our study also illustrates that estimates from a conditional odds ratio can be much larger than estimates from a relative risk ratio. Specifically, the conditional odds ratio from matched designs consider baseline risks and distinguish excess counts (not explained by the matching) from total counts (partially shared by the matching). An analysis based on excess counts examines only discordant cases; therefore, the conditional odds ratio reflects only the cases with differing outcomes.

An analysis based on total counts includes all the cases including those with congruent outcomes; therefore, the relative risk estimate reflects many cases with identical outcomes. These two estimates will differ most when the matching is informative and the baseline is substantial.

The yearly count of motorcycle fatalities increased substantially during our study due to many factors reflected in growing vehicle ownership (true also in many other countries).[30] Motorcycles are a practical means of transportation with a modest purchase price, efficient fuel consumption, low carbon emissions, and high mechanical reliability compared to the average car.[31] The compact size can also reduce traffic congestion and access easy parking.

For enthusiasts, no alternative offers the same exhilaration, recreational tourism, and supportive community.[32] Motorcyclists are also sometimes connected to political lobbying, social action, and renegade defiance.[33]

All these points highlight community values and serve as reminders that statistical science must respect culture and preferences.

Newsmedia Backlash

Since publishing our main findings in the *British Medical Journal*, we have been interviewed by many journalists.[34] The most common question is "How did you get this idea?" In our roles as a physician at Canada's largest trauma center (who has treated many patients injured in life-threatening motorcycle crashes) and the director of Princeton's Center for Behavioral Science and Public Policy (who has extensively studied attention and cognitive limitations), we noticed the frequent appearance of the full moon in advertising and the arts. This led to conversations about how small distractions such as the full moon might be one of a few causes of inattention that can be studied using large-scale rigorous science, due to the predictability of celestial mechanics.

A related question is why no one, to our knowledge, previously considered that a full moon could distract motorcyclists and lead to deaths. One explanation is that the observed association is small and hard to appreciate with casual observation. An additional factor is that individuals in crashes, even if they survive, can rarely describe what was on their minds before impact because fleeting distractions often fail to reach conscious awareness and a subsequent concussion will erase all memory of the precipitating event. Indeed, a patient's typical reply when asked "What happened?" is often "I don't know, I was knocked out."

Faulty introspection and deficient self-report, therefore, make it difficult to learn about motorcycle risks even from those who experience a minor motorcycle crash.

We could think of several explanations of how attention and perception might contribute to an association between the full moon and motorcycle fatalities. A full moon is infrequent and beautiful, and thus a natural distraction.[35] A full moon may contribute to increased outdoor activity of all types, including more-frequent journeys, longer distances, unfamiliar destinations, heightened cross-traffic, and mixtures of less-experienced travelers.[36]

A full moon can highlight other distractions that otherwise would be hidden in the dark.[37] It causes enhanced illumination of surrounding landscapes that may lead riders to misgauge distances and misjudge speeds (similar to aerial perspective bias for mountaineers, aviators, and astronauts in bright sunlight).[38]

Some journalists were motorcycle enthusiasts and raised further ideas about other vulnerabilities

[30]Bureau of Transportation Statistics. 2009. Motorcycle trends in the United States. Washington: U.S. Department of Transportation [DOT SR 014]. *http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/special_reports_and_issue_briefs/special_report/2009_05_14/pdf/entire.pdf.*

[31]Yperman, I. 2011. Commuting by motorcycle: impact analysis. *Transport and Mobility*. Leuven, Belgium.

[32]Horvath, P., Zuckerman, M. 1993. Sensation seeking, risk appraisal, and risky behaviour. *Personality and Individual Differences, 14*:41-52.

[33]Bosmia, A.N., Quinn, J.F., Peterson, T.B., Griessenauer, C.J., Tubbs, R.S. 2014. Outlaw motorcycle gangs: aspects of the one-percenter culture for emergency department personnel to consider. *Western Journal of Emergency Medicine, 15*(4):523–8.

[34]Redelmeier, D.A., Shafir, E. 2017. The full moon and motorcycle related mortality: population-based double control study. *British Medical Journal, 359*:j5367.

[35]Wilson, F.A., Stimpson, J.P. 2010. Trends in fatalities from distracted driving in the United States, 1999 to 2008. *American Journal of Public Health, 100*(11):2,213–9.

[36]Evans, L. 2004. *Traffic safety*. Bloomfield Hills, MI: Science Serving Society.

[37]Hershenson, M. 1989. *The moon illusion*. Hillsdale, NJ: Lawrence Erlbaum Associates.

[38]Redelmeier, D.A., Raza, S. 2018. Optical illusions and life-threatening traffic crashes: A perspective on aerial perspective. *Medical Hypotheses, 114*:23–27.

when riding a motorcycle unrelated to the full moon. Unlike driving a car, riding a motorcycle provides only a minimal base of support, occupant protection, roadway traction, visibility in traffic, and crash protection. Motorcycles also lack engineering advances such as seatbelts, airbags, protective interior engineering, a collapsible steering column, electronic stability control, adaptive cruise-control, and automatic blind-spot detectors. A motorcycle rider also has remarkable—and potentially distracting—sightlines due to the absence of a compartment roof.

One more distinction, of course, is that a self-driving motorcycle is unlikely to be feasible or popular anytime in the foreseeable future.

No journalist asked about statistical calculations or analytic techniques. In particular, the complexities of matched designs with conditional analyses did not arise even once in mainstream media conversations. We believe this indicates a greater (not lesser) duty on statisticians to understand nuances such as matched analyses, the interpretation of conditional odds ratio estimates, and the distinction between total risk and excess risk.

We also continue to encourage more reporting of summary counts for transparency and for basic calculations before hefty statistical modeling. Medical scientists, safety regulators, policy makers, and the general public hold high esteem for statisticians, so misinterpretations might otherwise easily arise and go without question.

One dismissive reaction from journalists was to trivialize the observed association as just one death for every two full moons. However, a relative risk of this magnitude exceeds the safety gains from reflective clothing or daytime running lights that are standard for motorcycle riders.

More generally, motorcycle crashes are not confined to a single country so, even without adjusting for differing road safety conditions, the global burden might extrapolate to roughly 10 deaths from the full moon each month worldwide. Of course, the moon is remote, inert, and predictable, so the combined effects of other distractors might be much greater. We hope that motorcyclists will appreciate the fundamental general point that small distractions can have major consequences and thus will ride with more care than dictated by intuition.

A final incendiary question was whether motorcycle riding is more dangerous than drunk driving. When expressed as a risk of death statistic, the typical ride on a motorcycle is twice as dangerous as a drunk driver traveling the same distance (35 vs. 15 deaths per 100 million vehicle-miles, respectively). Overall, a dozen adults get onto motorcycles during an ordinary day in the United States and do not survive; however, the life-time risk of a fatal crash is still less than 20% for a regular motorcyclist. Throughout, riders cannot sense the risks until it is too late (unlike, for example, hunger, which triggers eating well ahead of life-threatening starvation).

One purpose of science is to provide evidence so individuals might better appreciate the ongoing balance between freedom and safety.

Further Reading

Fatality Analysis Reporting System Encyclopedia. National Highway Traffic Safety Administration. *http://www-fars.nhtsa.dot.gov/Main/index.aspx*. Accessed February 12, 2015.

National Highway Traffic Safety Administration. 2012. Traffic Safety Facts. *http://www-nrd.nhtsa.dot.gov/Pubs/812032.pdf*. Accessed February 12, 2015.

Redelmeier, D.A., Shafir, E. 2017. The full moon and motorcycle related mortality: population based double control study. *British Medical Journal, 359*:j5367.

Redelmeier, D.A., Tibshirani, R.J. 2017. A simple method for analyzing matched designs with double controls: An extension of McNemar's test. *Journal of Clinical Epidemiology, 81*:51–55.

Redelmeier, D.A., Tibshirani, R.J. Methods for analyzing matched designs with double controls: excess risk is easily estimated and misinterpreted. *Journal of Clinical Epidemiology* (forthcoming).

About the Authors

Donald Redelmeier is a professor of medicine at the University of Toronto and holds the Canada Research Chair in Medical Decision Sciences. He earned his medical degree from the University of Toronto, completed residency and fellowship training at Stanford University Hospital, and received a master's degree in health services research from Stanford University. He practices medicine at Canada's largest trauma center and directly cares for patients when mistakes in reasoning lead to life-threatening consequences.

Eldar Shafir is the Class of 1987 Professor of Behavioral Science and Public Policy at Princeton University; inaugural director of Princeton's Kahneman-Treisman Center for Behavioral Science and Public Policy; and co-founder and scientific director at ideas42, a social science R&D lab. He studies decision-making, cognitive science, and behavioral economics. His recent research has focused on decision-making in contexts of poverty and on the application of behavioral research to policy. He is the co-author of *Scarcity: Why Having Too Little Means So Much.*

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

Deborah V. Dawson and Derek R. Blanchette

Scientific investigation does not march to its conclusions on a straight and open road. Studies may be inconclusive, or even produce conflicting results. They may have defects in design or execution that leave their conclusions open to question. The body of completed studies addressing a particular question may vary considerably in critical aspects such as design, population, measurement, or clinical/experimental protocols, all of which may cloud the interpretation of results and make meaningful comparisons difficult. This process is further complicated by the sheer magnitude of the ever-expanding scientific literature.

Readers frequently turn to review articles to obtain an overview of the available literature, and to attempt to get a grasp on the level of consensus—and controversy—associated with a particular line of scientific inquiry. Over the last four decades, biomedical research has increasingly turned to the systematic review format to synthesize and provide an overview of the available primary research pertinent to a particular scientific question. Such reviews are associated with careful delineation of a research question, followed by development of a detailed research protocol before commencing the review, including explicit specification of inclusion criteria.

This is followed by a comprehensive search to identify relevant studies that meet those criteria, and the application of rigorous methods for collection and evaluation of the available information, resulting in the research synthesis. Such procedures have been extensively codified by bodies such as the Cochrane Collaboration. The goal is to limit bias, and thus improve the reliability and accuracy of conclusions.

The intention is also that the comprehensive systematic review approach will help researchers avoid some of the pitfalls of the more-traditional "narrative" type of review. While traditional-style reviews have undoubtedly been of value, particularly when constructed by skillful, knowledgeable authors with a balanced outlook, they may suffer from inadequate literature searches and may be inefficient when reviewing a large number of studies. Further, by their very nature, they may involve selection bias—the reviewers' choices may reflect personal biases, focus on the best (or worst!) studies, or tend to highlight conflicts rather than resolve them.

In contrast, the systematic review aims to provide a comprehensive representation of all of the relevant literature. In some instances, a systematic review is accompanied by a quantitative synthesis of the results of the identified studies; this statistical treatment is termed a *meta-analysis*.

What Is Meta-Analysis?

The term meta-analysis originated in 1975 with Seymour Glass, whose efforts at research synthesis focused on psychotherapy outcomes. The "meta" does not refer to change (as in *metamorphosis*), but is more in the spirit of transcendence (as in *metaphysics*)—a going beyond analysis, producing a study of studies. When researchers seek to summarize the available literature through a systematic review with a quantitative synthesis, this constitutes a meta-analysis. Therefore, the rigorous search and synthesis protocols of the systematic review are considered part and parcel of the meta-analysis.

The late Ingram Olkin, who made significant contributions to meta-analysis, was known to say that the statistical portion of most meta-analyses was quite straightforward and almost minor—that the real effort was to be thorough and meticulous in the initial and critical tasks of identifying the question, developing protocols, and completing the search and abstraction activities.

Why Meta-analysis?

What can the reader reasonably expect to gain from digesting a

completed meta-analysis? As mentioned, meta-analysis can assist in answering a specific scientific question by summarizing data from multiple studies. It may be, for example, that multiple small studies are each inconclusive, but when combined via a meta-analytic synthesis, can provide a definitive answer. Similarly, individual studies may not have a sample size large enough to compare results in subgroups based on sex or ethnicity, but this may be possible when the information in these studies is combined via meta-analysis.

Meta-analysis can be helpful when studies disagree, by showing where the weight of the evidence lies, and can help resesarchers explore the nature of inconsistencies. In fact, a particularly useful feature of meta-analysis is its ability to examine and explore the heterogeneity in study outcomes—it may provide some useful information about the sources of variability among study findings.

Finally, meta-analysis can help identify gaps and problems in the published literature, and provide information for the planning of future, more-definitive, studies.

Examining Heterogeneity in Meta-analysis

Evaluation of variability is at the heart of meta-analysis. The very interpretation of the results of meta-analysis hinges on how heterogeneity among those studies is conceptualized. On the one hand, if all of the studies are measuring the same thing—the difference in functional gain between the particular treatment and placebo—and there is one right answer, one number corresponds to that average difference. Then, studies may vary in their estimate of that one right answer, but that variation is basically random noise. The combined estimate from the meta-analysis is thus an estimate of that one fixed number.

This conceptualization is referred to as the *fixed effects model*. But what if the mean difference itself is truly different in some studies? What if the impact of treatment really does differ for men and women, for older or younger subjects, or for delivery by one method or another?

In such cases, we would have a *distribution* of mean treatment differences in the component studies, and the combined estimate from the meta-analysis represents the mean difference in treatment effect in that population of studies. The variability associated with that estimate would have two components: the random noise component and the component associated with study-to-study variability.

Study-to-study variation might arise from differences in study protocols, populations, approaches to measurement, and so forth. This conceptualization is called the *random effects model*.

The meta-analytic method used will differ depending on the conceptual framework—which of these two situations is believed to reflect reality? It is possible to formally test whether there is significant study-to-study heterogeneity within the framework of meta-analysis.

If there is strong evidence of heterogeneity among eight component studies, it would be reasonable to adopt the random effects approach. However, often there are cogent reasons to suspect heterogeneity among the studies identified from the literature—for example, in surgical studies, different groups of investigators may use varying surgical protocols, or use different definitions of what constitutes a good outcome. In such instances, meta-analysts often elect to adopt a random effects model. It will be a somewhat more-conservative approach, and will tend to yield wider confidence intervals for the effect estimate. It also protects from using an inappropriate model where heterogeneity goes undetected.

Other explorations of heterogeneity are also possible with meta-analysis. It is possible to see which studies make the greatest contributions to heterogeneity. Identification of these studies may provide some interesting information, and potentially clues to the source of the heterogeneity.

For example, suppose that in a meta-analysis of dental treatments, one particular study provides the largest contribution to heterogeneity by far. Upon closer inspection, it is discovered that this study included molars and premolars, while the others were confined to only a single tooth type—the first molar. This would suggest that treatment effects may differ with tooth type, representing a potential explanation for the observed heterogeneity.

It is also possible to formally compare subgroups of studies to see whether the effect estimates differ significantly among subgroups of studies defined by surgical protocol, population type, and so forth. Such explorations can provide useful suggestions for further research.

Combining and Displaying the Results of Individual Studies

After formulating the research question, the investigator formalizes the hypothesis to be tested and selects the appropriate statistical measure to reflect that hypothesis. The research then abstracts the estimate of this quantity from

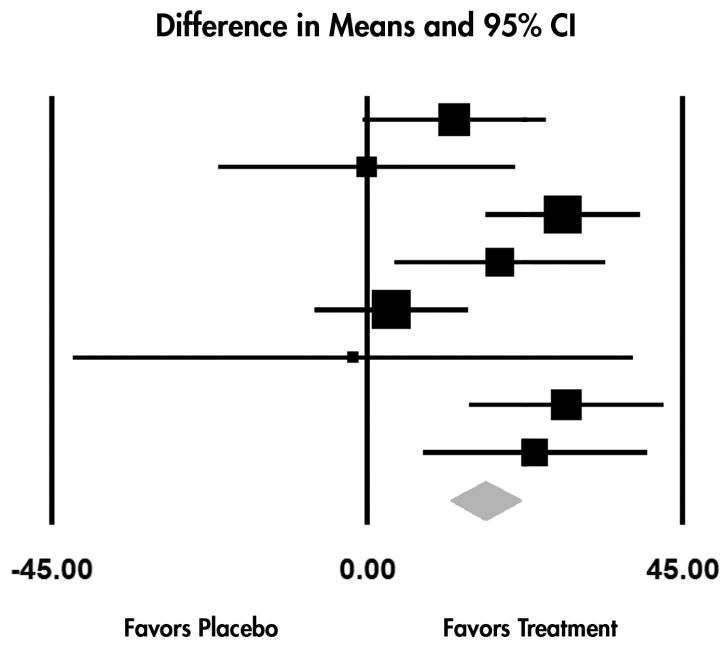


Figure 1. Synthetic data provide an example of a forest plot.
In this hypothetical example, the quantitative outcome is a score denoting gain of function after injury. Participants in each of the eight component studies were randomized to treatment or placebo. The effect estimate of interest is the mean difference in gain of function between treatment and placebo, and the value zero corresponds to no mean difference in gain of function between treatment and control. The effect estimates of individual studies are represented by squares, and the bars denote the associated 95% confidence intervals (CI). The size of the square is inversely related to the standard error, and reflects the weight given to the study in the meta-analysis. The diamond shape at the bottom represents the combined estimate from all available data arrived at through meta-analysis. It indicates that on average, the gain in functional score is 16.9 points higher in subjects receiving treatment than in those receiving placebo. The forest plot was created using Comprehensive Meta-Analysis, version 2.2.064 (Biostat, Inc.).

each of the various studies identified through the search protocol.

For example, if the intent is to compare the effect of two treatments, as measured by a quantitative outcome, the mean difference between treatment responses is the effect estimate of interest. We would require the difference in sample averages, and its associated standard error, obtained as a measure of variability, for each of the component studies. Other types of investigations will use other summary measures.

Regardless of whether the effect estimate of interest is the difference between treatment responses, a measure of risk (such as an odds ratio), or a measure of association (such as a correlation coefficient), these quantities and their associated standard errors must be identified accurately and recorded from the extant literature. They are then combined via meta-analytic techniques to obtain an overall estimate of the effect, as well as an estimate of the variability of this combined estimate, a pooled standard error.

In the absence of heterogeneity, a fixed effects model may be used; when there is evidence of heterogeneity in effect sizes among studies, or if there is reason to suspect heterogeneity, it calls for using the methods associated with the random effects model. However, it may be noted that, in practice, power to test for heterogeneity may be limited, providing further argument for the use of the random effects model. If the data suggest little heterogeneity, the random effects model would be expected to give results not too different from those of the fixed effects model in the majority of cases. The growing emphasis on the random effects approach in the literature may reflect these considerations.

Whichever model is used, the results can be used to obtain confidence intervals, and to formally test hypotheses. The combined results are often presented along with the results from the individual component studies in a graphical display known as a *forest plot*.

Figure 1 provides such a plot. In this example, based on synthetic data, the quantitative outcome

is a measure of gain in function after injury. In each of the eight component studies, there are two independent groups of subjects—those randomized to the treatment group and those randomized to receive a placebo. The effect estimate of interest is the mean difference in outcome (gain in function) between treatment and placebo, which is calculated as (treatment mean—placebo mean) and for each study, its estimate is denoted by a square. A value of zero corresponds to no effect of treatment relative to placebo.

For each of the eight studies in Figure 1, the effect estimate is provided along with a 95% confidence interval, indicated by a horizontal bar. The interpretation of a such a confidence interval is that there is 95% confidence that the true value of the mean difference in treatment effect for that study has been captured within the limits of the interval.

We notice that the mean differences for the different studies vary considerably. In this example, the majority of the studies show results favoring treatment, but one favors placebo, however slightly; another shows that, on average, there is no difference between treatment and placebo. In some instances, the 95% confidence interval for a particular study straddles the zero line, implying there is no evidence of a statistically significant difference between treatment and placebo based upon that particular study. The widths of the confidence intervals also vary considerably, reflecting the precision associated with each study.

Larger studies tend to have considerably greater precision and thus narrower confidence intervals, compared to smaller studies.

Finally, the size of the squares denoting the effect estimate (the mean difference in functional gain between treatment and placebo) varies among the studies; the size of the square is inversely proportional to the standard error—studies with greater precision (smaller standard errors) have larger squares and more weight in the meta-analysis. Studies with less precision (greater standard errors) have smaller squares and are weighted less heavily.

If the formal test of study-to-study variability is performed for the example given in Figure 1, there is strong evidence of heterogeneity among the eight component studies, and the random effects approach is used to estimate the combined effect size and its standard error. At the bottom of the figure, the diamond shape conveys information about the overall estimate of the mean difference in gain in function between treatment and placebo. The diamond is centered at the overall effect estimate and its breadth reflects the corresponding meta-analytic 95% confidence interval, which are both derived from the combined information from all available studies, weighted as just described, based upon the random effects model.

In this instance, we see that patients receiving treatment had greater gain in function than those on placebo—an average of 16.9 points more. This is our estimate, based upon all available data of the average effect size among this population of studies. Its associated standard error includes components representing both random variation and study-to-study variability. The resulting 95% confidence interval for the combined estimate is 11.9 to 22.0. This interval does *not* include the value zero that corresponds to the case where there is no difference between treatment and placebo.

The meta-analysis can also provide a *P*-value associated with the formal statistical test for treatment effect, again based on the combined data from the eight studies. In this case, $P < 0.0001$, and this very small value provides strong evidence that the treatment is associated with greater gain in function than the placebo. The forest plot has provided insight into the variation among the component studies, as well as a representation of the combined results of the meta-analysis, which strongly supports the superiority of treatment relative to placebo.

Meta-analysis of Individual Participant-level Data

Another context in which meta-analytic techniques may be applied is when a meta-analysis is conducted at the level of the individual participant data, rather than based upon the aggregate results taken from the component studies. These aggregate results were, of course, derived from the individual participant data, so, in a sense, the meta-analyst is going back to the original source material.

The goal of the meta-analysis remains the same: to answer a particular question by summarizing the available evidence. However, the use of the raw, individual level data in the meta-analysis may permit adjustment for confounding factors, and may increase the power to detect differential treatment effects for subgroups—provided that the necessary information is available in the component studies. It may provide a way to impose consistency in inclusion/exclusion criteria across all studies, and provides the opportunity to assess the validity of model assumptions associated with statistical methods.

Further, meta-analysis at the individual level allows the analyst to use approaches that take account of the correlation among multiple outcomes, such as when subjects are followed longitudinally. (It is crucial in such meta-analyses that the within-study clustering of the data is recognized and retained in the conduct of the analysis; it would be quite inappropriate to simply pool data from all individuals and treat them as if they came from a single study.)

Addressing Concerns about Bias

The possibility of publication bias is a serious threat to the validity of a meta-analysis. Studies showing no effect of treatment, or no difference between treatments, are less likely to be published, particularly when such "negative" studies have limited sample sizes. Even when such studies do come to publication, it may be in the form of a shorter, less-detailed communication, which may or may not contain sufficient information to permit inclusion in a meta-analysis.

How can one claim to represent the whole of the research evidence that has been produced if a proportion of that evidence never gets to publication? A number of methods, both analytic and graphical (e.g., funnel plots), have been developed to try to identify publication bias. It may, however, be difficult to apply them when the number of publications in the relevant literature is limited.

Meta-analysts also strive to address the potential problem of publication bias by searching for relevant abstracts, scrupulously including articles in any language, and directly contacting authors to try to fill in gaps in the published literature. They may also look for relevant information in the so-called "gray literature"—literature outside the traditional publishing and distribution channels. Examples might include theses and dissertations, technical reports, unpublished studies, or studies published outside broadly available mainstream journals.

It is similarly important to consider the possibility of other types of bias, including bias in the component studies contributing to the meta-analysis. Suboptimal conduct of research may lead to bias—for example, because of differential attrition among treatment groups, biased selection of subjects, inadequate randomization protocols, or not masking subjects or evaluators with respect to which treatment was assigned. The quality of the component studies is critical to the validity of the meta-analysis.

The Movement Toward Quality Assessment

In recent years, there has been an increasing emphasis on quality assessment—in terms of both the quality of the meta-analysis itself, and the quality of the component studies. Both have bearing on the validity and usefulness of the meta-analysis. A number of checklists intended to address quality of the meta-analysis have been developed for use by authors, editors, reviewers, decision-makers—and readers!

These checklists are worth consulting by the reader of the scientific literature, since they provide a useful guide to the characteristics possessed by a sound and well-conducted meta-analysis.

Typically, a checklist will include items to help assess the quality and rigor associated with the different phases of the meta-analysis—the search protocol, selection of inclusion criteria, abstraction procedures, analytic approach, and efforts to avoid publication bias. One such checklist is Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA), which has 27 items targeting aspects ranging from the formulation of the research question through the methodology and synthesis of results.

Other checklists are specifically tailored to assist in evaluating either meta-analyses of clinical trials, such as QUOROM (Quality of Reporting of Meta-analyses), or those intended for use with observational studies, such as Meta-analysis Of Observational Studies in Epidemiology (MOOSE).

Equally important is the quality of the component studies. For this reason, increasing scrutiny has been brought to bear on the conduct of these contributing studies. In the case of a clinical trial, appropriate questions to ask include: Was the study randomized? Was the person assessing the clinical outcome blinded about the treatment applied? Were there any conflicts of interest?

One instrument that tries to look at the potential impact of such features in a systemic manner is the Cochrane Collaboration tool for assessing risk of bias in randomized trials. These considerations are paramount, since one can hardly expect a meta-analysis, however carefully and conscientiously conducted, to overcome the deficiencies and biases of the studies that contribute to it.

Consensus?

Not everyone is entirely enamored of meta-analysis. Titles in the literature addressing meta-analysis include phrases such as "statistical alchemy" and even "meta-analysis/shmeta-analysis." Clearly,

as discussed, the meta-analyst cannot expect to be a magician, transcending all shortcomings of the studies in the extant literature.

Nevertheless, provided that the potential pitfalls of meta-analysis—such as poorly conducted component studies and the risk of publication bias—are recognized, meta-analysis can be trusted and useful. It can provide a quantitative synthesis of the existing literature with a measure of objectivity. In many cases, being armed with the additional power of its combined sample sizes will not only provide consensus and resolve conflicts in the published literature, but may even make it possible to answer questions that cannot be addressed by any of the single studies, such as inferences related to subgroups, or dose-response effects.

For all of these reasons, meta-analysis has become a fundamental tool in efforts to attain evidence-based clinical practice. Finally, even in relative failure, where consensus is not reached or the painstaking methodology of meta-analysis reveals a less-than-pretty state of the current science, meta-analysis has its contributions to make: It makes us cognizant of the failings of the available evidence, whether through deficiencies in sample size, design or execution. Where evidence is scant and there are gaps in the literature, it provides suggestions for next steps.

About the Authors

Deborah V. Dawson is a professor of biostatistics in the Iowa Institute for Oral Health Research, and the first recipient of the Morris Bernstein Professorship in Dentistry. She served for 16 years as director of biostatistics and research design at the University of Iowa College of Dentistry. Dawson received her ScM in biostatistics from Johns Hopkins University and her PhD in biostatistics from the University of North Carolina at Chapel Hill. Her research interests focus on dental caries and enamel defects, gene-environment interaction, meta-analysis, and the history of statistics—especially the origins of the discipline of biometry.

Derek R. Blanchette is a statistician in the Division of Biostatistics and Computational Biology at the University of Iowa College of Dentistry. He received his MS in biostatistics from the University of Iowa College of Public Health and is accredited by the American Statistical Association. His interests include statistical computing, data visualization, and meta-analysis.

Further Reading

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

Brusselaers, N. 2015. How to teach the fundamentals of meta-analyses, *Annals of Epidemiology, 25*:948–54.

Dawson, D.V., Pihlstrom, B.L., and Blanchette, D.R. 2016. "Understanding and evaluating meta-analysis," *Journal of the American Dental Association, 147*:264–70.

Sutton, A.J., Abrams, K.R. Jones, D.R. Sheldon, T.A., and Song, F. 2000. *Methods for Meta-Analysis in Medical Research*. Chichester: Wiley & Sons Ltd.

Dalene Stangl and Mine Çetinkaya-Rundel
Column Editors

Top 10 Reasons Not to Drop Insignificant Regressors: A Statistical Listicle

Chris Rohlfs

Picture this: You design an ordinary least squares (OLS) regression model and run it. Based on the results of that first model, you drop the statistically insignificant regressors and re-run. You repeat multiple times until you finally arrive at a model in which all the regressors (or the ones you care about) are statistically significant.

You can find strategies like this being used all over the place. Stepwise regression, an automated variant of this approach first proposed in 1960, returns more than half a million hits on Google, including lecture materials from university courses and a number of instructional websites, and the phrase appears in tens of thousands of academic articles published in the past decade. The Wikipedia page for stepwise regression includes a criticism section, but is on balance neutral to positive, and functions for different versions of stepwise estimation exist in all of the major statistical packages. Beyond the use of specific automated techniques, evidence from the literature indicates that selective presentation of regressions with statistically significant coefficients is widespread in many areas of academia.

Despite its ubiquity in practice, those who are most knowledgeable about the subject discourage researchers from using this approach to regression. The American Statistical Association recently published a statement recommending against using explicit thresholds for statistical significance. Many statisticians and economists criticize the overemphasis on statistical significance and the misapplication of the concept in discussing and designing regression models; in the top echelons of these disciplines, automated stepwise regression is rarely used or even discussed.

The aim of this article is to explain, for people who aren't statistics or economics professors, why regression analysis of this form is misguided. Here are the top 10 reasons not to drop insignificant regressors.

P-hacking is the process of cherry-picking models to obtain statistically significant results. The arguments against *p*-hacking as presented here apply to automated approaches such as stepwise regression and to informal trial-and-error. Even when researchers do not admit to p-hacking, its effects can be seen through publication bias, where statistically significant findings are more common in academic journals than would be expected due to random chance.

10. Those Standard Errors Don't Mean What You Think They Do

Whether you're *p*-hacking or not, the standard errors produced by OLS are a little unreliable. They only give correct measures of the variability in your estimates if your residuals are independent and identically distributed, but that is rarely the case. It is possible to use modified standard error formulas that relax those assumptions; however, these require their own sometimes-restrictive assumptions about how different observations' residuals are correlated with one another. If you use a formula that is based upon the wrong set of assumptions, then your estimated standard errors will be incorrect, and you will be wrong about which regressors are statistically significant.

Luckily, you can put more of your trust in the coefficients themselves. Your residuals can all have different distributions with different shapes, and any of them can be correlated positively or negatively with any others. As long as the relationship you are estimating is linear and there is no omitted

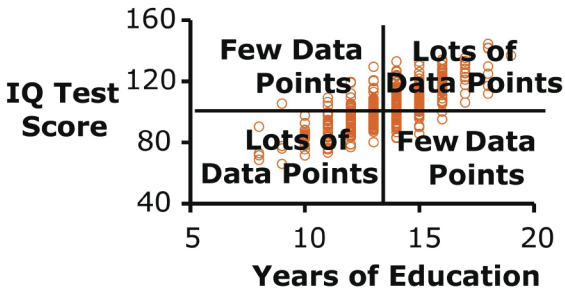


Figure 1. Sample size by IQ and education (hypothetical data).

Common alternatives to the default OLS standard errors are robust or White standard errors that relax the identical distribution assumption, Newey-West standard errors that allow for correlation between nearby observations in a time-series, and clustered standard errors that allow you to specify groups of observations that might be correlated with one another (for instance, survey respondents from the same household).

variable bias, your coefficient estimates will be right on average across repeated samples. If your model selection process is guided by significance tests, which rely upon the standard errors, you lose this robustness, and your estimated coefficients—and even your specification itself—will have the same sensitivity to ancillary model assumptions that plagues your standard errors.

9. It's Just Plain Wrong

A standard error is supposed to say how much your coefficient estimates are likely to move if you re-run your statistical procedure on an entirely new sample. If you have been running regression models, changing the specifications, and re-running, the standard errors from the final regression specification tell you how much your estimates would move if you re-ran that final estimation step on a new sample, ignoring all the testing that you did to get that specification. However, things could have gone wrong in a lot more steps than just the final estimation.

Those results are more sensitive to the data—and are more likely to move around from one sample to another—than you are admitting. As empirical researchers, we always run the risk that the relationships we identify exist only in our sample and not generally; we use standard errors and significance tests to gauge how likely that event is. If the standard errors, t-statistics, and p-values that you present ignore the model search process, then you are overstating your level of confidence in your findings. Do it too often, and you will be accused of crying wolf.

Wait, you say. How do you know I didn't just come up with the significant regressors from the start? What if I'm just really lucky or know how to pick winners? Why would a lucky person get to take credit for those statistically significant coefficients, but I don't?

The problem with this logic is that, lucky or not, a solid study relies on credible assumptions. By leaving out certain potential predictors, you are assuming *a priori* that they have zero effect on your outcome variable. You might have justifications for excluding them, but the stories you are telling cannot be too convincing. If they were, you would not have wasted your own time considering those candidate predictors in your model search process.

8. Get Your Priorities Straight

A common mistake in data work is to confuse statistical significance with materiality or economic importance. Statistical significance depends upon both the magnitude of the estimated coefficient and the level of precision with which it is estimated. Regressors sometimes have large but statistically insignificant coefficients, so your best guess is that the variable is an important predictor—possibly the most important determinant of the outcome variable—but you are uncertain about whether that is true.

This problem is magnified when you have two or more important regressors that are correlated with one another, as illustrated in Figure 1. Suppose that we are running a regression to predict workers' earnings (not shown), and two of the predictors being considered are IQ test scores and years of education. Most of the people in our data who achieved high levels of education also have high scores on the IQ test, and most of the people with lower levels of education perform poorly on the IQ test.

When you run a regression that includes both variables as predictors, the estimates of each variable's

contribution are driven primarily by the people in the data who have unusually high or low IQ scores relative to their levels of education, as in the upper-left and lower-right boxes of Figure 1.

The sample sizes in these segments of the population are small, so we are likely to arrive at imprecise estimates of the effects of both IQ and education. We might know that at least one of the two predictors is important, but cannot say with confidence that it is IQ or that it is education. As a result, the coefficients on both variables could be statistically insignificant.

If you design your regression specification by filtering out statistically insignificant predictors, you could end up dropping both of these potentially important contributors to earnings, or you could pick one as the key driver based upon the arbitrary "horse race" of statistical significance—and you would end up overstating your level of confidence in that decision.

It is well-known that models with correlated regressors can produce volatile and unstable coefficient estimates, but if, *a priori*, you did not have an economic rationale for dropping one or more of those predictors, then this volatility reflects true uncertainty about the appropriate model. That uncertainty is correctly signaled to the reader in the form of large standard errors on your coefficients, supposing that you've specified your standard errors in the right way (see reason 10).

7. You Are Not Helping

What do you have to gain by dropping insignificant regressors? Not much, really. They are not causing any bias in your other estimates. They are not hurting anyone. In fact, those insignificant coefficient estimates provide information about what variables are not effective at explaining movements in the outcome variable or have effects that are hard to measure precisely. For the sake of parsimony, you might exclude a regressor that you believe *a priori* to be unimportant. You might also exclude a regressor if you do not want it held constant as other regressors in your model vary.

Selecting an appropriate set of predictors requires applying domain knowledge and common sense, and that is your primary task as a modeler, but if you are not already excluding a regressor for those reasons, obtaining a statistically insignificant coefficient is no reason to drop it.

6. Slow Down There, Buddy

If you are trying to choose what to eat for breakfast or what clothes to wear in the morning, it helps to be decisive, but if you are studying a question that has implications for medicine, safety, public policy, or business, you want to convey the risks correctly. One of the most valuable assets you can have as a researcher is a reputation as someone who uncovers general truths about the world—and one of the greatest threats to this reputation arises when another researcher tries to replicate your work and discovers that your findings are sensitive to slight variations in the way you set up your model.

To demonstrate how confident readers should feel about your forecasts, you can perform this replication yourself and explore the robustness of your results to different specifications. There is more than just one way to think about whatever problem you are studying. If you want to tell the whole story, show us how your results change when you use different sensible sets of predictors.

Perfect multi-collinearity arises when one of the regressors is an exact function of the others, so you cannot estimate all the coefficients in the regression equation. Partial multi-collinearity arises when one of the regressors is a partial function of the others, so you can produce estimates, but the standard errors are larger than you might have expected, because it is hard to figure out which predictors are the more important ones. There is no real solution to partial multi-collinearity, except to recognize that the estimates would be more precise if you had more data (which is always true).

You have to make all sorts of judgment calls when you perform a statistical study, such as what data set to use and what outcome variables and predictors to include in your model. In some cases, the decisions are fairly arbitrary; in others, reasonable people might disagree about the best way to do it. If you experiment with lots of different combinations of these judgment calls and they all lead to the same general statistical result, then that result is robust.

5. ... It's complicated

One of the advantages of OLS regression is that it is simple and can be performed with a single line of code or a few mouse clicks. If you drop insignificant regressors and re-run, you add a layer of complication to this process and another step where something could go wrong and where critical readers can raise concerns—and you make your research less transparent.

That process also involves judgment calls. You could opt to use an automated stepwise regression procedure, in which case you would have to select among forward selection, backward elimination, bidirectional elimination, or other

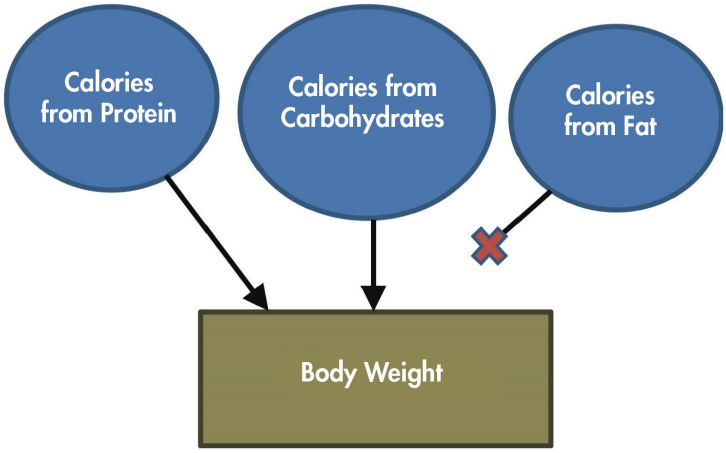


Figure 2. Incomplete model of body weight.

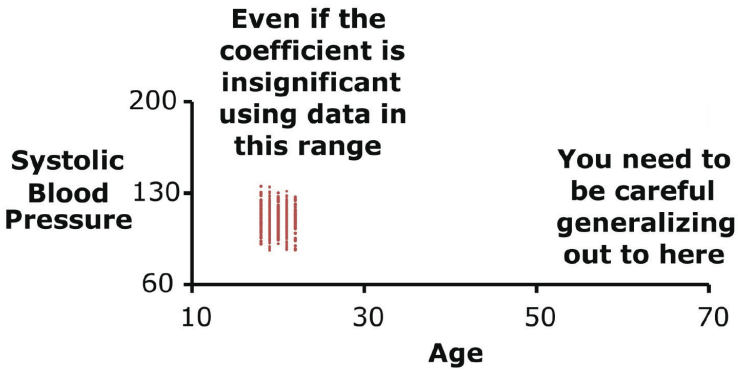


Figure 3. Generalizing effect of age on blood pressure (hypothetical data).

approaches. You also might choose the informal approach of running regressions, checking the results, and modifying the specifications in whatever way you like.

This toxic combination of a lack of transparency and the need for judgment calls means that readers will have trouble understanding what you did and why. The more jaded among us will suspect you of using your discretion to twist the data into producing a certain result—and then using the complication of the procedure to hide it.

4. A False Sense of Simplicity

Occam's razor suggests that we should favor simple models over more-complicated ones. Having a small number of predictors, however, does not necessarily make a model simple. What makes a model simple is that its assumptions are obvious—they are basically the same as ones we are already using in other contexts, so the model does not require us to rethink how the world works.

Figure 2, for instance, illustrates a model that would be simplified by adding another parameter. Suppose that you wanted to estimate how much calorie consumption causes people to gain weight, and your potential predictors include calories from protein, carbohydrates, and fat.

A simple model would say that the type of calorie should not matter, so we would expect similar-sized positive coefficients on all three predictors. If you were dropping statistically insignificant regressors, you could arrive at the model shown in Figure 2, which excludes one of the three types of calories.

Although this new model includes fewer predictors, it is more complicated in the sense that it is counterintuitive. It would be a deviation from how we think about the way the body works, and anyone looking at the model would ask why you left out the third type of calorie. If you construct an initial regression specification with a large number of potential predictors, that model could be simple or complicated, depending upon the assumptions under which you derive the specification. Dropping some of those regressors based upon statistical significance may make your model look simpler, but the choice of predictors will no longer be guided by intuition.

As a result, you might have to resort to convoluted explanations for why some predictors are included in the model while other equally defensible ones are not.

3. Think Outside the Data

The reason you develop a model and test it with data is to derive general principles about biological systems, the economy, or whatever application you are considering. If you use a data-driven model selection approach such as dropping insignificant regressors, then you

are likely to arrive at an entirely different final specification when you try to replicate your process on a new sample. If you cannot rely on your model to be the same when you change your data set, how can you use it with confidence on real world applications?

To illustrate one type of problem with this approach, consider the example in Figure 3, which considers age as a predictor of blood pressure. If the sample only includes college students, we are likely to obtain a statistically insignificant coefficient, because the differences in age in the sample are too small for their effects to be estimated precisely. Within our sample, there may be little direct effect of including or ignoring age as a predictor. However, what if we want to generate out-of-sample predictions, and what if the population we want to apply our results to includes people in their 50s and 60s?

Someone who was simply mining the data might look at the clump of observations in Figure 3 (or the statistical results they would generate) and rule out age as a possible predictor. A careful researcher, though, who was taking domain knowledge into account would recognize the shortcomings of such out-of-sample estimation—that a predictor that might not look important in the original sample could be very important when generalizing to a different population.

Reasonable solutions to this problem might be to collect new data or to explain to the reader how data limitations make it difficult to obtain precise estimates. One of the worst solutions to this problem is to leave age out of the model entirely, effectively ignoring its effects and claiming with confidence that they are zero.

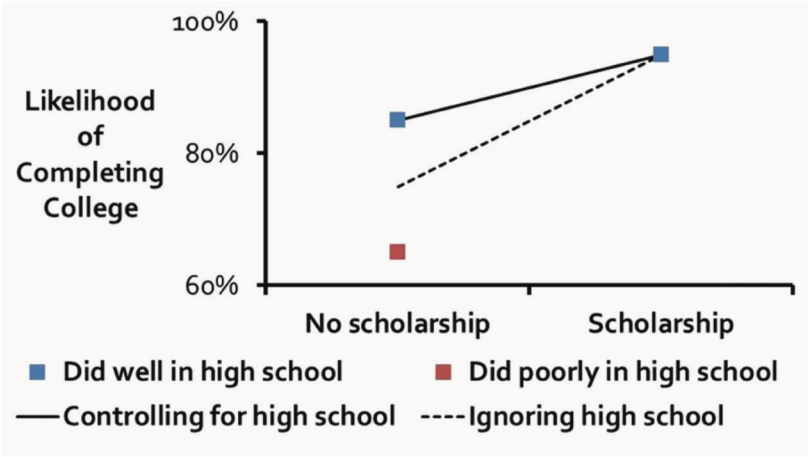


Figure 4. Scholarships and college completion (hypothetical data).

An increasingly popular way to make estimates more generalizable is to separate the data set into a training sample on which the coefficients or other model parameters are estimated, a validation sample that is used to judge which of the different estimated models has greater forecasting accuracy, and a test sample that is used to assess the predictive power of the final selected model. While the formal properties of this approach are not known, it has been applied with great success in the field of machine learning.

Omitted variables bias arises when one of the regressors correlates with an important predictor that you left out of the equation. When this happens, instead of estimating the true causal effect of your regressor, your estimates provide some combination of the effect of the regressor and the effect of the confounding variable. You can solve this problem by including the confounder as a control variable in the regression or by using an experiment or natural experiment.

2. Be a Control Freak

Some regressors are less important for their direct effects on the outcome variable than for the ways they interact with other predictors. A classic example is control variables. Consider the example shown in Figure 4, where we are trying estimate the effects of receiving a scholarship on the likelihood of finishing college. Suppose that scholarships are given to a fraction of the students who did well in high school and to none of the students who did poorly in high school. If we only measured the difference in college completion rates between scholarship recipients and non-recipients, our

estimates would suffer from omitted variables bias.

That is, we would measure some mixture of the effect of receiving the scholarship and the effect of doing well in high school, as shown by the slope of the dashed line. If we want to estimate the true causal effect of the scholarship, then we need to control for the effects of high school performance, and we would arrive at the slope of the solid line. Even if the coefficient on high school performance (our control variable) is statistically insignificant, leaving it out of our regression specification would bias our estimate of the effect of the scholarship.

About the Author

Chris Rohlfs is an executive director of wealth management strategies and modeling at Morgan Stanley, where he has worked for the past five years. He has published on a variety of applied topics, from national defense to demography to natural language processing. Before working at Morgan Stanley, he was an assistant professor of economics in the Maxwell School at Syracuse University for seven years. He received his PhD in economics from the University of Chicago in 2006 and his MS in computer science from New York University in 2015, and is currently working part-time on a doctorate of engineering sciences in electrical engineering at Columbia University, studying deep learning.

1. The 1970s Called ...

That's right: The 1970s called, and they want their statistical procedure back.

Since the 1970s, there have been enormous advances in the ways that researchers think about data and regression. Computer scientists and statisticians have developed machine learning approaches to replicate complex systems for classification and prediction, as well as methods for working with enormous data sets. In the economics world, researchers have refined practices for measuring causal relationships, dedicating greater resources to randomized trials and devising new techniques for simulating experimental environments with observational data.

This means that dropping insignificant regressors and re-running your model—whether through an informal iterative approach or using stepwise regression—takes time and energy that you could spend on these other promising avenues of research or simply finding better data, improving the research design by learning more details about your data set and the factors that cause the regressors to vary across observations, or performing sensitivity testing on the model.

Further Reading

American Statistical Association. 2016. ASA statement on statistical significance and *p*-values. *The American Statistician*, 70: 131–133.

Angrist, J.D., and Pischke, J.-S. 2009. *Mostly harmless econometrics: an empiricist's companion*. Princeton, NJ: Princeton University Press.

Gelman, A. 2014. Statistical modeling, causal inference, and social science: Why we hate stepwise regression (blog entry). Retrieved from *http://andrewgelman.com/2014/06/02/hate-stepwise-regression/*.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Sribney, B. 2005. Problems with stepwise regression. Stata Resources and Support: FAQs. Retrieved from *http://www.stata.com/support/faqs/statistics/stepwise-regressionproblems/*.

Ziliak, S.T., and McCloskey, D.N. 2008. The cult of statistical significance: how the standard error costs us jobs, justice, and lives. Ann Arbor, MI: University of Michigan Press.

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON



READY FOR A CHANGE?

Transform your data career with a flexible master's degree or certificate from UW–Madison.

Data careers are exciting, important, and lucrative. And the **demand** for savvy data wranglers continues to grow. Build on your current **skills, knowledge, and experience**. No matter which aspect of data interests you most—**analytics to visualization**, or something in between, UW–Madison has a path for you.

Explore 13 data science and analytics programs:

- **Business**
  - Capstone Certificate in Actuarial Science
  - Capstone Certificate in Data Analytics for Decision Making
  - Master of Science in Economics
  - Master of Science in Statistics
- **Computer Science**
  - Capstone Certificate in Computer Sciences
  - Master of Science in Computer Sciences
- **Engineering**
  - Master of Engineering in Engineering
- **Environment/Sustainability**
  - Master of Science in Agricultural and Applied Economics
  - Master of Science in Environmental Conservation
- **GIS**
  - Capstone Certificate in GIS Fundamentals
  - Capstone Certificate in Advanced GIS
  - Master of Science in Cartography and Geographic Information Systems

We offer programs with flexible formats that **fit the lives of working adults**. A degree or certificate from UW–Madison will **advance your career**.

Visit **go.wisc.edu/exploreuwdata**

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.