Vol. 32, No. 1, 2019

CHANCE

Using Data to Advance Science, Education, and Society

Including...

Big Data and Privacy

How We Know that the Earth is Warming

09332480(2019)32(1)

Taylor & Francis
Taylor & Francis Group

ASA

EXCLUSIVE BENEFITS
FOR ALL ASA MEMBERS!

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

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

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

DATA VISUALIZATION
Charts, Maps, and Interactive Graphs

Robert Grant, BayersCamp

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

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

VISUALIZING BASEBALL

Jim Albert, Bowling Green State University, Ohio, USA

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

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

ERRORS, BLUNDERS, AND LIES
How to Tell the Difference

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

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

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

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
Vol 112, 2017

THE AMERICAN STATISTICIAN
Vol 72, 2018

STATISTICS AND PUBLIC POLICY
Vol 5, 2018

Taylor & Francis Group
an **informa** business

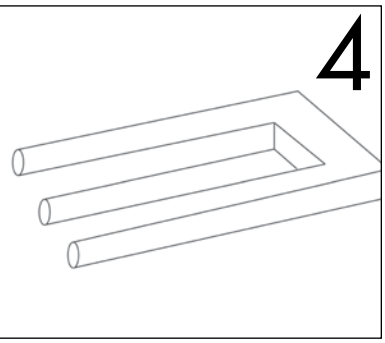
ASA

http://bit.ly/CRCASA2018

CHANCE

Using Data to Advance Science, Education, and Society

http://chance.amstat.org



ARTICLES

COLUMNS

DEPARTMENTS

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

Cover image: Melissa Gotherman

AIMS AND SCOPE

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

SUBSCRIPTION INFORMATION

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

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

ASA MEMBER SUBSCRIPTION RATES

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

SUBSCRIPTION OFFICES

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

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

OFFICE OF PUBLICATION

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

RESPONSIBLE FOR ADVERTISEMENTS

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

EDITOR'S LETTER

Scott Evans

Dear CHANCE *Colleagues,*

Happy new year! To bring in this new year, we are playing some of *CHANCE*'s greatest hits by revisiting some of the most-popular articles through the years and mixing in a few new ones.

Our first article was quite thought-provoking and one of the forerunners of today's growing discussion regarding the use of *p*-values. **John Ioannidis** discusses why most published research findings are false!

Tatsuki Koyama then discusses how the length of the Beatles' songs increased during the latter part of their career. (On another Beatles topic, check out **Mark Glickman** on ScienceFriday: *https://www.sciencefriday.com/segments/who-wrote-that-beatles-song-this-algorithm-will-tell-you/*. John Lennon claimed to have written "In My Life," but Paul McCartney remembers it differently. Mark evaluates who is telling the truth.)

Browsing further through the *CHANCE* photo album, **Kevin Bales** unlocks statistics from a special issue of *CHANCE* on modern slavery. Derek Young evaluates 007…after all he informs us that "James Bond will return." **Peter Guttorp** explains how we know the Earth is warming in his article from a special issue of *CHANCE* on climate change.

Let's move to today: Data science is all the rage! **Miguel Hernan**, **John Hsu**, and **Brian Healy** discuss using data science to redefine data analysis to accommodate causal inference from observational data. They offer a classification of data science tasks.

Tom Adams then shares an excerpt from his exciting new book, *Improving your NCAA Bracket with Statistics*.

We also re-visit one of our column articles from yesteryear: **Nicole Lazar** discusses "Big Data and Privacy" from the Big Picture column. Nicole wraps up the article with, "To all of my students, former students, collaborators past and present, and old friends who try to connect via LinkedIn, Facebook, ResearchGate, and the like—when I don't respond, just know that I don't participate in any of those fora. It's my small way of keeping a corner of privacy in the world." I bet Nicole is good at filling out NCAA brackets.

We end this issue with **Christian Robert** reviewing *Pragmatics of Uncertainty* by Joseph Kadane; *10 Great Ideas About Chance* by Persi Diaconis and Brian Skyrms; *Independent Random Sampling Methods* by Luca Martino, David Luengo, and Joaquin Miguez; and *Computational Methods for Analysis with R* by James Howard.

We hope that you enjoy the trip down memory lane.

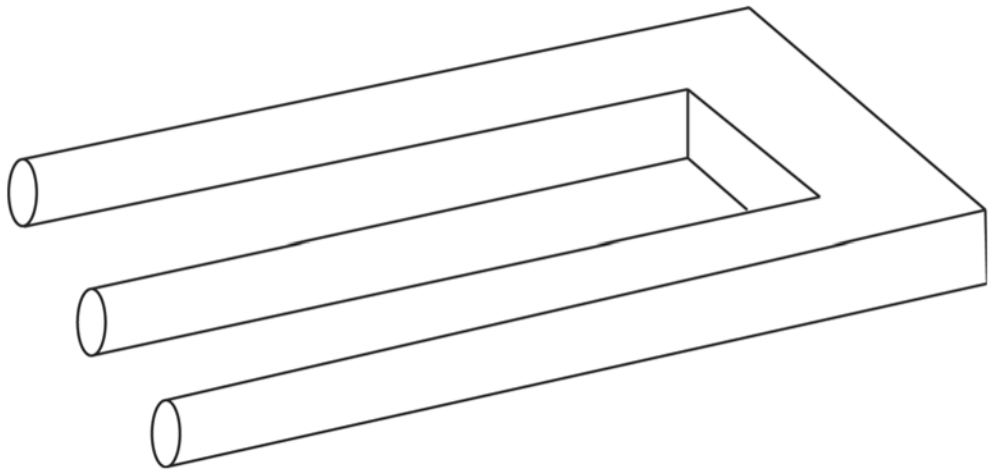
Scott Evans

Correction: In the *CHANCE* 31.4 article "Elementary Statistics on Trial—The Case of Lucia de Berk" by Richard D. Gill, Piet Groeneboom, and Peter de Jong, the caption for the Fokke Sukke cartoon on page 11 refers to "two ducks." It should be: "The canary and the duck are defending a family guardian…"

Why Most Published Research Findings Are False

John P.A. Ioannidis

Reprinted courtesy of the Public Library of Science



Editor's note: This article contains opinion on topics of broad interest and does not necessarily reflect the views of CHANCE.

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field.

In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; effect sizes are smaller; there is a greater number and lesser preselection of tested relationships; there is greater flexibility in designs, definitions, outcomes, and analytical modes; there is greater financial and other interest and prejudice; and more teams are involved in a scientific field in cases of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings often may be simply accurate measures of the prevailing bias.

Published research findings are refuted sometimes by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies to the most modern molecular research. There is increasing concern that in modern research, false findings may be the majority, or even the vast majority, of published research claims. However, this should not be surprising. It can be proven that most claimed research findings are false.

This article originally appeared in *CHANCE* 18.4.

Modeling the Framework for False-Positive Findings

Several methodologists have pointed out that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded, strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented

Table 1—Research Findings and True Relationships

Research Finding	True Relationship		
	Yes	No	Total
Yes	$c(1-\beta)R/(R+1)$	$c\alpha/(R+1)$	$c(R+\alpha-\beta R)/(R+1)$
No	$c\beta R/(R+1)$	$c(1-\alpha)/(R+1)$	$c(1-\alpha+\beta R)/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c

DOI: 10.1371/journal.pmed.0020124.t001

and summarized by *p*-values, but unfortunately, there is a widespread notion that medical research articles should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance (e.g., effective interventions, informative predictors, risk factors, or associations). 'Negative' research also is very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance. Consider a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field, both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of true relationships to no relationships among those tested in the field. R is characteristic of the field and can vary a lot, depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated.

Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or there is roughly equal power for finding any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate).

The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value (PPV).

The PPV is also the complementary probability of what has been called the false positive report probability. According to the 2 × 2 table, one gets $PPV = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus more likely true than false if $(1 - \beta)R > \alpha$. Because usually the vast majority of investigators depend on $\alpha = 0.05$, a research finding is more likely true than false if $(1 - \beta)R > 0.05$.

What is less appreciated is that bias and the extent of repeated independent testing by different teams of investigators around the globe may further distort this picture and may lead to even smaller probabilities of the research findings being indeed true. We will try to model these two factors in the context of similar 2 × 2 tables.

Bias

First, let us define bias as the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced. Let u be the proportion of probed analyses that would not have been 'research findings,' but nevertheless end up presented and reported as such because of bias.

Bias should not be confused with chance variability that causes some findings to be false by chance, even though the study design, data, analysis, and presentation are perfect. Bias can entail manipulation in the analysis or reporting of findings. Selective or distorted reporting is a typical form of such bias. We may assume that u does not depend on whether a true relationship exists. This is not an unreasonable assumption, as

Table 2—Research Findings and True Relationships in the Presence of Bias

Research Finding	True Relationship		
	Yes	No	Total
Yes	$(c[1-\beta]R + uc\beta R)/(R+1)$	$(c\alpha + uc(1-\alpha))/(R+1)$	$c(R+\alpha-\beta R+u-u\alpha+u\beta R)/(R+1)$
No	$(1-u)c\beta R/(R+1)$	$(1-u)c(1-\alpha)/(R+1)$	$c(1-u)(1-\alpha+\beta R)/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c

DOI: 10.1371/journal.pmed.0020124.t002

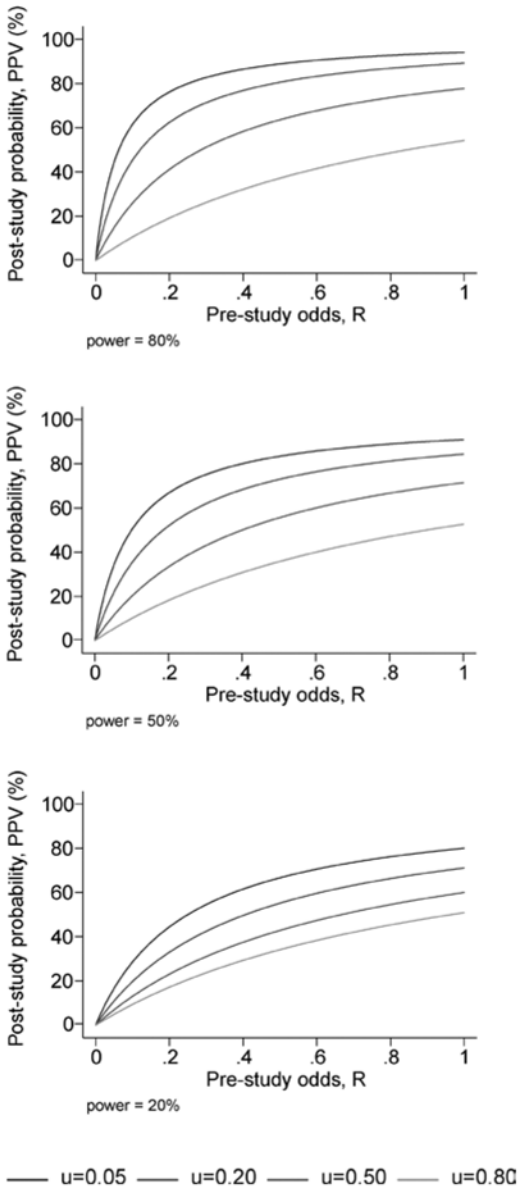


Figure 1. PPV (Probability that a Research Finding is True) as a Function of the Pre-study Odds for Various Levels of Bias, u (panels correspond to power of 0.20, 0.50, and 0.80).

typically it is impossible to know which relationships are indeed true.

In the presence of bias (Table 2), one gets PPV = $([1-\beta]R + u\beta R)/(R + \alpha - \beta R + u - u\alpha + u\beta R)$, and PPV decreases with increasing u, unless $1-\beta \leq \alpha$, (i.e., $1-\beta \leq 0.05$), for most situations. Thus, with increasing bias, the chance that research findings are true diminishes considerably. This is shown for different levels of power and for different pre-study odds in Figure 1.

Conversely, true research findings may occasionally be annulled because of reverse bias. For example, with large measurement errors, relationships are lost in noise, or investigators use data inefficiently or fail to notice statistically significant relationships, or there may be conflicts of interest that tend to "bury" significant findings.

There is no good large-scale empirical evidence of how frequently such reverse bias may occur across diverse research fields. However, it is probably fair to say that reverse bias is not as common. Moreover, measurement errors and inefficient use of data are probably becoming less-frequent problems, since measurement error has decreased with technological advances in the molecular era and investigators are becoming increasingly sophisticated about their data.

Table 3—Research Findings and True Relationships in the Presence of Multiple Studies

Research Finding	True Relationship		
	Yes	No	Total
Yes	$cR(1-\beta^n)/(R+1)$	$c(1-[1-\alpha]^n)/(R+1)$	$c(R+1-[1-\alpha]^n-R\beta^n)/(R+1)$
No	$cR\beta^n/(R+1)$	$c(1-\alpha)^n/(R+1)$	$c([1-\alpha]^n+R\beta^n)/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c

DOI: 10.1371/journal.pmed.0020124.t003

Regardless, reverse bias may be modeled in the same way as bias above. Also, reverse bias should not be confused with chance variability that may lead to missing a true relationship because of chance.

Testing by Several Independent Teams

Several independent teams may be addressing the same sets of research questions. As research efforts are globalized, it is practically the rule that several research teams, often dozens of them, may probe the same or similar questions. Unfortunately, in some areas, the prevailing mentality until now has been to focus on isolated discoveries by single teams and interpret research experiments in isolation. An increasing number of questions has at least one study claiming a research finding, and this receives unilateral attention. The probability that at least one study, among several on the same question, claims a statistically significant research finding is easy to estimate. For n independent studies of equal power, the 2 × 2 table is shown in Table 3: PPV = $R(1-\beta^n)/(R+1-[1-\alpha]^n-R\beta^n)$ (not considering bias).

With an increasing number of independent studies, PPV tends to decrease, unless $1-\beta<\alpha$ (i.e., typically $1-\beta<0.05$). This is shown for different levels of power

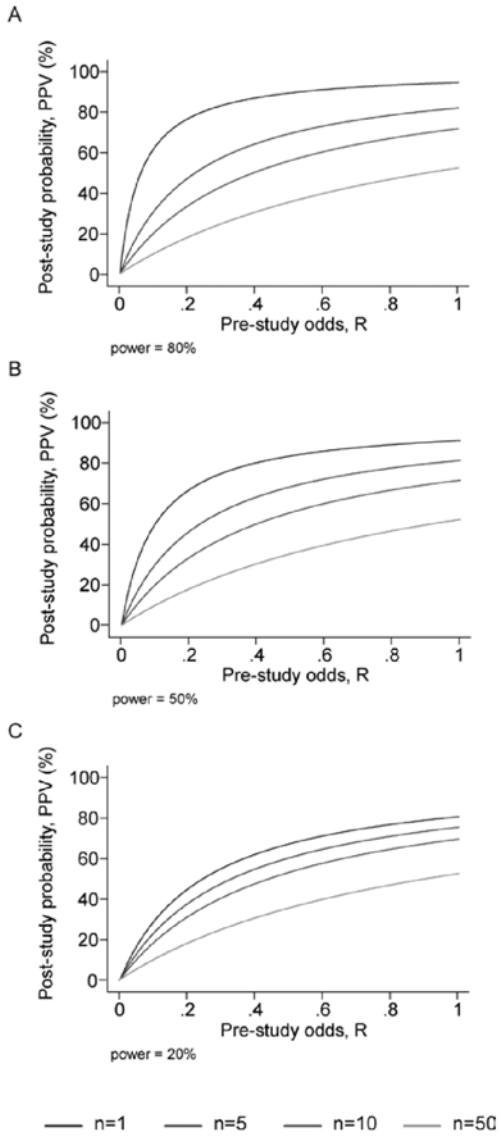


Figure 2. PPV (Probability that a Research Finding is True) as a Function of the Pre-study Odds forVarious Numbers of Conducted Studies, n (panels correspond to power of 0.20, 0.50, and 0.80).

and for different pre-study odds in Figure 2. For n studies of different power, the term β^n is replaced by the product of the terms βi for i = 1 to n, but inferences are similar.

Corollaries

Based on the above considerations, one may deduce several interesting corollaries about the probability that a research finding is indeed true.

Corollary 1: The smaller the studies conducted in a scientific field, the less likely the research findings are to be true. Small sample size means smaller power and, for all functions above, the PPV for a true research finding decreases as power decreases toward 1 – β = 0.05. Thus, other factors being equal, research findings are more likely true in scien- tific fields that undertake large studies, such as randomized controlled trials in cardiology (several thousand subjects randomized), than in scientific fields with small studies, such as most research of molecular predictors (sample sizes 100-fold smaller).

Corollary 2: The smaller the effect sizes in a scientific field, the less likely the research findings are to be true. Power also is related to the effect size. Thus, research findings are more likely true in scientific fields with large effects, such as the impact of smoking on cancer or cardiovascular disease (relative risks 3–20), than in scientific fields where postulated effects are small, such as genetic risk factors for multigenetic diseases (relative risks 1.1–1.5). Modern epidemiology is increasingly obliged to target smaller effect sizes. Consequently, the proportion of true research findings is expected to decrease. In the same line of thinking, if the true effect sizes are very small in a scientific field, this field is likely to be plagued by almost-ubiquitous false-positive claims. For example, if the majority of true genetic or nutritional determinants of complex diseases confer relative risks of less than 1.05, genetic or nutritional epidemiology would be largely utopian endeavors.

Corollary 3: The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true. As shown above, the post-study probability that a finding is true (PPV) depends a lot on the pre-study odds (R). Thus, research findings are more likely true in confirmatory designs, such as large phase III randomized controlled trials, or meta-analyses thereof, than in hypothesis-generating experiments. Fields considered highly informative and creative given the wealth of the assembled and tested information, such as microarrays and other high-throughput discovery-oriented research, should have extremely low PPV.

Corollary 4: The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true. Flexibility increases the potential for transforming what would be 'negative' results into 'positive' results (i.e., bias, u). For several research designs (e.g., randomized controlled trials or meta- analyses), there have been efforts to standardize their conduct and reporting. Adherence to common standards is likely to increase the proportion of true findings. The same applies to outcomes. True findings may be more common when outcomes are unequivocal and universally agreed (e.g., death), rather than when multifarious outcomes are devised (e.g., scales for schizophrenia outcomes). Similarly, fields that use commonly agreed, stereotyped analytical methods (e.g., Kaplan-Meier plots and the log-rank test) may yield a larger proportion of true findings than fields where analytical methods are still under experimentation (e.g., artificial intelligencemethods) andonly "best" results are reported. Regardless, even in the most stringent research designs, bias seems to be a major problem. For example, there is strong evidence that selective outcome reporting, with manipulation of the outcomes and analyses reported, is a common problem even for randomized trails. Simply abolishing selective publication would not make this problem go away.

Corollary 5: The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true. Conflicts of interest and prejudice may increase bias (u). Conflicts of interest are very common in biomedical research, and typically they are inadequately and sparsely reported. Prejudice may not necessarily have financial roots. Scientists in a given field may be prejudiced purely because of their belief in a scientific theory or commitment to their own findings. Many otherwise seemingly independent, university-based studies may be conducted for no other reason than to give physicians and researchers qualifications for promotion or tenure. Such nonfinancial conflicts also may lead to distorted reported results and interpretations. Prestigious investigators may suppress, via the peer review process, the appearance and dissemination of findings that refute their findings, thus condemning their field to perpetuate false dogma. Empirical evidence on expert opinion shows it is extremely unreliable.

Corollary 6: The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true. This seemingly paradoxical corollary follows because, as stated

Table 4—PPV of Research Findings for Various Combinations of Power (1 − β), Ratio of True to Not-true Relationships (R), and Bias (u)

1 − β	*R*	*u*	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

The estimated PPVs (positive predictive values) are derived assuming $\alpha = 0.05$ for a single study.
RCT, randomized controlled trial.
DOI: 10.1371/journal.pmed.0020124.t004

above, the PPV of isolated findings decreases when many teams of investigators are involved in the same field. This may explain why we occasionally see major excitement followed rapidly by severe disappointments in fields that draw wide attention. With many teams working in the same field and with massive experimental data being produced, timing is of the essence in beating competition. Thus, each team may prioritize on pursuing and disseminating its most impressive "positive" results. "Negative" results may become attractive for disseminationonly if some other team has found a "positive"association on the same question.

In that case, itmay beattractive to refute a claim made in some prestigious journal. The term "Proteus phenomenon" has been coined to describe rapidly alternating extreme research claims and equally extreme opposite refutations. Empirical evidence suggests this sequence of extreme opposites is very common in molecular genetics.

These corollaries consider each factor separately, but these factors often influence each other. For example, investigators working in fields where true effect sizes are perceived to be small may be more likely to perform large studies than investigators working in fields where true effect sizes are perceived to be large. Prejudice may prevail in a hot scientific field, further undermining the predictive value of its research findings.

Highly prejudiced stakeholders may even create a barrier that aborts efforts at obtaining and disseminating opposing results. Conversely, the fact that a field is hot or has strong invested interests may sometimes promote larger studies and improved standards of research, enhancing the predictive value of its research findings. Massive discovery-oriented testing also may result in such a large yield of significant relationships that investigators have enough to report and search further and thus refrain from data dredging and manipulation.

Most Research Findings Are False for Most Research Designs and for Most Fields

In the described framework, a PPV exceeding 50% is quite difficult to get. Table 4 provides the results of simulations using the formulas developed for the influence of

power, ratio of true to non-true relationships, and bias for various types of situations that may be characteristic of specific study designs and settings. A finding from a well-conducted, adequately powered, randomized controlled trial, starting with a 50% pre-study chance that the intervention is effective, is eventually true about 85% of the time.

A fairly similar performance is expected of a confirmatory meta-analysis of good-quality randomized trials: Potential bias probably increases, but power and pre-test chances are higher compared to a single randomized trial. Conversely, a meta-analytic finding from inconclusive studies where pooling is used to "correct" the low power of single studies, is probably false if $R \leq 1{:}3$.

Research findings from underpowered, early-phase clinical trials would be true about one in four times, or even less frequently if bias is present. Epidemiological studies of an exploratory nature perform even worse, especially when underpowered, but even well-powered epidemiological studies may have only a one-in-five chance of being true, if $R = 1{:}10$.

Finally, in discovery-oriented research with massive testing, where tested relationships exceed true ones 1,000-fold (e.g., 30,000 genes tested, of which 30 may be the true culprits), PPV for each claimed relationship is extremely low, even with considerable standardization of laboratory and statistical methods, outcomes, and reporting to minimize bias.

Claimed Research Findings Often May Simply Be Accurate Measures of the Prevailing Bias

As shown, the majority of modern bio-medical research is operating in areas with very low pre- and post-study probability for true findings. Let us suppose that in a research field there are no true findings at all to be discovered. The history of science teaches us that scientific endeavor has wasted effort in fields with absolutely no yield of true scientific information often in the past, at least based on our current understanding. In such a null field, one would ideally expect all observed effect sizes to vary by chance around the null in the absence of bias. The extent that observed findings deviate from what is expected by chance alone would be simply a pure measure of the prevailing bias.

For example, let us suppose that no nutrients or dietary patterns are actually important determinants for the risk of developing a specific tumor. Let us also suppose that the scientific literature has examined 60 nutrients and claims all of them to be related to the risk of developing this tumor with relative risks in the range of 1.2 to 1.4 for the comparison of the upper to lower intake tertiles. Then, the claimed effect sizes are simply measuring nothing but the net bias that has been involved in the generation of this scientific literature.

Claimed effect sizes are in fact the most accurate estimates of the net bias. It even follows that between null fields, the fields that claim stronger effects (often with accompanying claims of medical or public health importance) are simply those that have sustained the worst biases.

For fields with very low PPV, the few true relationships would not distort this overall picture much. Even if a few relationships are true, the shape of the distribution of the observed effects would still yield a clear measure of the biases involved in the field.

This concept totally reverses the way we view scientific results. Traditionally, investigators have viewed large and highly significant effects with excitement, as signs of important discoveries. Too large and too highly significant effects actually may be more likely to be signs of large bias in most fields of modern research. They should lead investigators to careful critical thinking about what might have gone wrong with their data, analyses, and results.

Of course, investigators working in any field are likely to resist accepting that the whole field in which they have spent their careers is a null field. However, other lines of evidence, or advances in technology and experimentation, eventually may lead to the dismantling of a scientific field. Obtaining measures of the net bias in one field also may be useful for obtaining insight into what might be the range of bias operating in other fields where similar analytical methods, technologies, and conflicts are operating.

How Can We Improve the Situation?

Is it unavoidable that most research findings are false, or can we improve the situation? A major problem is that it is impossible to know with 100% certainty what the truth is in any research question. In this regard, the pure "gold" standard is unattainable. However, there are several approaches to improve the post-study probability.

Better-powered evidence (e.g., large studies or low-bias meta-analyses) may help, since it comes closer to the unknown gold standard. However, large studies still may have biases, and these should be acknowledged and avoided.

Moreover, large-scale evidence is impossible to obtain for all of the millions and trillions of research questions posed in current research.

Large-scale evidence should be targeted for research questions where the pre-study probability is already considerably high, so a significant research finding will lead to a post-test probability that would be considered quite definitive.

Large-scale evidence also is indicated, particularly when it can test major concepts rather than narrow, specific questions. A negative finding can then refute not only a specific proposed claim, but a whole field or considerable portion thereof. Selecting the performance of large-scale studies based on narrow-minded criteria, such as the marketing promotion of a specific drug, is largely wasted research. Moreover, one should be cautious that extremely large studies may be more likely to find a formally statistical significant difference for a trivial effect that is not meaningfully different from the null.

Second, most research questions are addressed by many teams, and it is misleading to emphasize the statistically significant findings of any single team. What matters is the totality of the evidence. Diminishing bias through enhanced research standards and curtailing of prejudices also may help. However, this may require a change in scientific mentality that might be difficult to achieve.

In some research designs, efforts also may be more successful with upfront registration of studies (e.g., randomized trials). Registration would pose a challenge for hypothesis-generating research. Some kind of registration or networking of data collections or investigators within fields may be more feasible than registration of each and every hypothesis-generating experiment.

Regardless, even if we do not see a great deal of progress with registration of studies in other fields, the principles of developing and adhering to a protocol could be more widely borrowed from randomized controlled trials.

Finally, instead of chasing statistical significance, we should improve our understanding of the range of R values—the pre-study odds—where research efforts operate. Before running an experiment, investigators should consider what they believe the chances are that they are testing a true, rather than a non-true, relationship. Speculated high R values then may be ascertained sometimes. As described above, whenever ethically acceptable, large studies with minimal bias should be performed on research findings that are considered relatively established to see how often they are indeed confirmed. I suspect several established "classic" studies will fail the test.

Nevertheless, most new discoveries will continue to stem from hypothesis-generating research with low or very low pre-study odds. We should acknowledge then that statistical significance testing in the report of a single study gives only a partial picture, without knowing how much testing has been done outside the

A WHOLE GENOME ASSOCIATION STUDY

Let us assume a team of investigators performs a whole genome association study to test whether any of 100,000 gene polymorphisms are associated with susceptibility to schizophrenia. Based on what we know about the extent of heritability of the disease, it is reasonable to expect that probably around 10 gene polymorphisms among those tested would be truly associated with schizophrenia, with relatively similar odds ratios around 1.3 for the 10 or so polymorphisms and with a fairly similar power to identify any of them. Then $R = 10/100{,}000 = 10^{-4}$, and the pre-study probability for any polymorphism to be associated with schizophrenia is also $R/(R + 1) = 10^{-4}$. Let us also suppose the study has 60% power to find an association with an odds ratio of 1.3 at $\alpha = 0.05$. Then it can be estimated that if a statistically significant association is found with the p-value barely crossing the 0.05 threshold, the post-study probability that this is true increases about twelve-fold, compared with the pre-study probability, but it is still only 12×10^{-4}.

Now let us suppose the investigators manipulate their design, analyses, and reporting so as to make more relationships cross the $p = 0.05$ threshold, even though this would not have been crossed with a perfectly adhered to design and analysis and with perfectcomprehensive reporting of the results, strictly according to the original study plan. Such manipulation could be done, for example, with serendipitous inclusion or exclusion of certain patients or controls, post hoc subgroup analyses, investigation of genetic contrasts thatwere notoriginally specified, changes in the disease or control definitions, and various combinations of selective or distorted reporting of the results. Commercially available data mining packages actually are proud of their ability to yield statistically significant results through data dredging. In the presence of bias with $u = 0.10$, the post-study probability that a research finding is true is only 4.4×10^{-4}. Furthermore, even in the absence of any bias, when 10 independent research teams perform similar experiments around the world, if one of them finds a formally statistically significant association, the probability that the research finding is true is only 1.5×10^{-4}, hardly any higher than the probability we had before any of this extensive research was undertaken!

report and in the relevant field at large. Despite a large statistical literature for multiple testing corrections, usually it is impossible to decipher how much data dredging by the reporting authors or other research teams has preceded a reported research finding. Even if determining this were feasible, this would not inform us about the pre-study odds. Thus, it is unavoidable that one should make approximate assumptions on how many relationships are expected to be true among those probed across the relevant research fields and research designs. The wider field may yield some guidance for estimating this probability for the isolated research project.

Experiences from biases detected in neighboring fields also would be useful to draw upon. Even though these assumptions would be considerably subjective, they would be useful in interpreting research claims and putting them in context. ©

John P.A. Ioannidis can be emailed at jioannid@cc.uoi.gr.

References

Altman, D.G., and Royston, P. 2000. What do we mean by validating a prognostic model? *Stat Med* 19:453–473.

Altman, D.G., and Goodman, S.N. 1994. Transfer of technology from statistical journals to the biomedical literature. Past trends and future predictions. *JAMA* 272:129–132.

Antman, E.M., Lau, J., Kupelnick, B., Mosteller, F., and Chalmers, T.C. 1992. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *JAMA* 268: 240–248.

Bartlett, M.S. 1957. A comment on D.V. Lindley's statistical paradox. *Biometrika* 44: 533–534.

Chan, A.W., Hrobjartsson, A., Haahr, M.T., Gotzsche, P.C., and Altman, D.G. 2004. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 291: 2457–2465.

Colhoun, H.M., McKeigue, P.M., and Davey Smith, G. 2003. Problems of reporting genetic associations with complex outcomes. *Lancet* 361:865–872.

De Angelis, C., Drazen, J.M., Frizelle, F.A., Haug, C., Hoey, J., et al. 2004. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N Engl J Med* 351:1250–1251.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531–537.

Hsueh, H.M., Chen, J.J., and Kodell, R.L. 2003. Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. *J Biopharm Stat* 13:675–689.

International Conference on Harmonisation E9 Expert Working Group 1999. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. *Stat Med* 18:1905–1942.

Ioannidis, J.P., and Trikalinos, T.A. 2005. Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. *J Clin Epidemiol* 58: 543–549.

Ioannidis, J.P., Evans, S.J., Gotzsche, P.C., O'Neill, R.T., Altman, D.G., et al. 2004. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med* 141:781–788.

Ioannidis, J.P. 2003. Genetic associations: False or true? *Trends Mol Med* 9:135–138.

Ioannidis, J.P., Haidich, A.B., and Lau, J. 2001. Any casualties in the clash of randomised and observational evidence? *BMJ* 322:879–880.

Ioannidis, J.P.A. 2005. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 294:218–228.

Ioannidis, J.P.A. 2005. Microarrays and molecular research: Noise discovery? *Lancet* 365:454–455.

Ioannidis, J.P.A., Ntzani, E.E., Trikalinos, T.A., and Contopoulos-Ioannidis, D.G. 2001. Replication validity of genetic association studies. *Nat Genet* 29:306–309.

Kelsey, J.L., Whittemore, A.S., Evans, A.S., and Thompson, W.D. 1996. *Methods in Observational Epidemiology*, 2nd ed. New York: Oxford University Press.

Krimsky, S., Rothenberg, L.S., Stott, P., and Kyle, G. 1998. Scientific journals and their authors' financial interests: a pilot study. *Psychother Psychosom* 67: 194–201.

Lawlor, D.A., Davey Smith, G., Kundu, D., Bruckdorfer, K.R., and Ebrahim, S. 2004. Those confounded vitamins: What can we learn from the differences between observational versus randomised trial evidence? *Lancet* 363:1724–1727.

Lindley, D.V. 1957. A statistical paradox. *Biometrika* 44: 187–192.

Marshall, M., Lockwood, A., Bradley, C., Adams, C., Joy, C., et al. 2000. Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. *Br J Psychiatry* 176:249–252.

Michiels, S., Koscielny, S., and Hill, C. 2005. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 365:488–492.

Moher, D., Schulz, K.F., and Altman, D.G. 2001. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 357:1191–1194.

Moher, D., Cook, D.J., Eastwood, S., Olkin, I., Rennie, D., et al. 1999. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of reporting of meta-analyses. *Lancet* 354:1896–1900.

Ntzani, E.E., and Ioannidis, J.P. 2003. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* 362:1439–1444.

Papanikolaou, G.N., Baltogianni, M.S., Contopoulos-Ioannidis, D.G., Haidich, A.B., Giannakakis, I.A. et al. 2001. Reporting of conflicts of interest in guidelines of preventive and therapeutic interventions. *BMC Med Res Methodol* 1:3.

Ransohoff, D.F. 2004. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer* 4:309–314.

Risch, N.J. 2000. Searching for genetic determinants in the new millennium. *Nature* 405: 847–856.

Senn, S.J. 2001. Two cheers for *p*-values. *J Epidemiol Biostat* 6:193–204.

Sterne, J.A., and Davey Smith, G. 2001. Sifting the evidence—What's wrong with significance tests? *BMJ* 322:226–231.

Stroup, D.F., Berlin, J.A., Morton, S.C., Olkin, I., Williamson, G.D., et al. 2000. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 283: 2008–2012.

Taubes, G. 1995. Epidemiology faces its limits. *Science* 269: 164–169.

Topol, E.J. 2004. Failing the public health—Rofecoxib, Merck, and the FDA. *N Engl J Med* 351: 1707– 1709.

Vandenbroucke, J.P. 2004. When are observational studies as credible as randomised trials? *Lancet* 363:1728–1731.

Wacholder, S., Chanock, S., Garcia-Closas, M., El ghormli, L., and Rothman, N. 2004. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 96:434–442.

Yusuf, S., Collins, R., and Peto, R. 1984. Why do we need some large, simple randomized trials? *Stat Med* 3:409–422.

Length of the Beatles' Songs

Tatsuki Koyama

"Length of the Beatles' Songs" shows the length and release date of each of the Beatles' 210 songs. Perhaps the figure should include more than 210 songs, since the list of Beatles songs on Wikipedia.org (visited September 5, 2011) gives 305 songs with the caveat, "many more…should be added." Two-hundred and ten is the number of songs released between the band's debut single and last studio album, give or take. Perhaps, "Kansas City/Hey-Hey-Hey-Hey" should be counted as two songs. Also, maybe "Revolution" (single version) and "Revolution 1" (album version) should be counted as two songs, given that the versions differ by nearly one minute in length.

This article originally appeared in *CHANCE* 25.1.

Song Length

Many criteria could be considered to illustrate The Beatles' transition from mop-headed pop idols to psychedelic cultural icons. I chose song length. Simple, straightforward, dry, statistical. To chronicle the Beatles' transition in any other way would probably require painstaking research and deep understanding of Sixties culture.

Beatles fans of all ages can probably agree that the band got weird as their career progressed. "Revolution 9" is weird, and perhaps so are "I Want You (She's So Heavy)" and "Her Majesty," but weirdness is subjective. Not everyone will agree which Beatles' songs are weird. By contrast, song length is objective and no cultural insight is required to make statements such as, "The Beatles' songs got longer as their career progressed." More importantly, song length data are readily available.

The data show something a bit more complex than "their songs got longer." The Beatles indeed had longer songs in the latter part of their career, but their shortest songs also came from that time. Dichotomization of a continuous variable is usually discouraged, but the distribution of the Beatles' song lengths, as visualized in the plot, begs for a dividing line right after "Revolver" (August 5, 1965) to split their career into early and late phases.

Most striking as seen in the plot is the consistency of the Beatles' song lengths in the early phase. The majority of the songs (110/128=86%) in this phase are between two and three minutes. This amazing consistency and complete lack of songs lasting more than four minutes (fairly common in the post-"Revolver" era) allow the main discussion of the plot to be presented in the empty space above the pre-"Revolver" data, enhancing the immediacy of the information describing the plot. Although obvious from a look, consistency of pre-"Revolver" song lengths is confirmed in the statistics, as shown in Table 1.

As well as providing the big picture of the song lengths, the plot also includes details (e.g., song

Table 1—Summary of Song Lengths on the Beatles' Studio Albums

Album	Release Date	Number of Songs	Song Length				
			Mean	SD	Min	Median	Max
Please Please Me	1963/3/22	14	2′20	0′21	1′50	2′23	2′57
With the Beatles	1963/11/22	14	2′23	0′20	1′48	2′22	3′02
A Hard Day's Night	1964/6/26	13	2′21	0′16	1′49	2′20	2′46
Beatles for Sale	1964/12/4	14	2′21	0′18	1′46	2′27	2′55
Help!	1965/8/6	14	2′24	0′20	1′54	2′23	3′10
Rubber Soul	1966/12/3	14	2′33	0′18	2′05	2′29	3′22
Revolver	1966/8/5	14	2′29	0′20	2′01	2′44	3′01
Sgt. Pepper	1967/6/1	13	3′05	1′06	1′20	2′54	5′33
The Beatles	1968/11/22	30	3′06	1′16	0′52	3′21	8′13
Yellow Submarine	1969/1/17	6*	3′37	1′23	2′10	2′45	6′28
Abbey Road	1969/9/26	17	2′46	1′38	0′23	3′20	7′47
Let It Be	1970/5/8	12	2′54	1′05	0′41	2′33	4′01

Table 2—Songs Composed by Rare Combinations of The Beatles

Title	Release Date	Credits	Length
What Goes On	1965/12/3	Lennon/McCartney/Starkey	2′50
Flying	1967/12/8	Lennon/McCartney/Harrison/Starkey	2′16
Don't Pass Me By	1968/11/22	Starkey	3′50
Octopus's Garden	1969/9/26	Starkey	2′51
Dig It	1970/5/8	Lennon/McCartney/Harrison/Starkey	0′49
Maggie Mae	1970/5/8	Lennon/McCartney/Harrison/Starkey	0′41

names) of interest. The unusually long and short songs are all labeled, as well as the songs mentioned in the discussion.

The composers are another piece of information depicted in the plot. Of 210 songs, 161 (77%) are credited to John Lennon-Paul McCartney and 20 to George Harrison. This leaves 23 cover songs and six compositional outliers (Table 2). "Maggie Mae," released on August 8, 1970, should probably be regarded as a cover song; however, all four Beatles were credited as composers of this public domain piece, robbing "Maggie" of the distinction of being the first cover song in nearly five years (since "Dizzy Miss Lizzy" and "Act Naturally," released in *Help!* on August 6, 1965) and the shortest song the Beatles ever covered. Crediting the Beatles for "Maggie Mae" maintains the truth of the claim: "All 23 cover songs came in the early phase."

Perhaps the Beatles are not the only band that shows these trends with regard to song lengths in this era. Superimposing similar data for another band or for the top 10 songs from each year might be of interest.

About the Author

Tatsuki Koyama is a research associate professor in the Division of Cancer Biostatistics, Department of Biostatistics, and an investigator at the Center for Quantitative Sciences at Vanderbilt University School of Medicine. His research has primarily focused on adaptive and flexible clinical trial design methodologies.

The Beatles released 210 songs, give or take—from their debut single, "Love Me Do" (released October 5, 1962) to their final studio album, *Let It Be* (May 8, 1970)—in just under eight years. This plot shows length versus release date of each song. All release dates are for the United Kingdom market with one exception ("Bad Boy").

Data and Notation

The closed circles indicate songs included on the Beatles' 12 studio albums, the titles and release dates of which appear in the abscissa. The Beatles also released 44 songs as singles, shown here with open circles. They released 16 songs twice each, both as singles and as part of an album; all these pairs are identical in both versions, except for "Revolution," "Get Back," and "Let It Be." "Yellow Submarine" was released three times—as a single and on *Revolver* (both on August 5, 1966) and on *Yellow Submarine* (January 17, 1969).

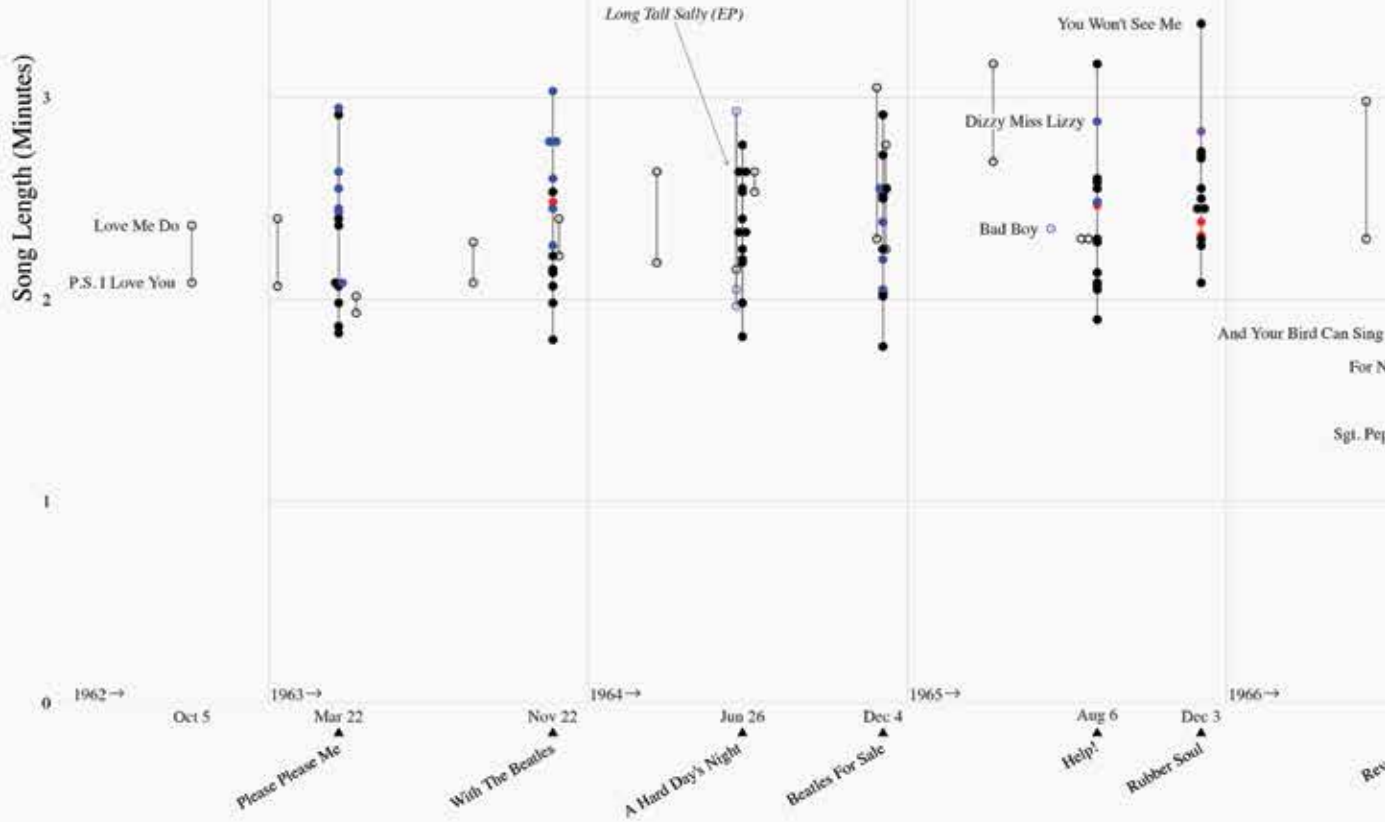
The studio albums and singles encompass all but 10 of the Beatles' 210 songs. Two of their 13 extended-play albums (EPs), *Long Tall Sally* and *Magical Mystery Tour*, consist of songs that were never released elsewhere. A total of nine songs from these two EPs also are shown in the plot. *Magical Mystery Tour* includes "I am the Walrus," which is not shown in the plot because it was previously released as a single. One song was never a part of any studio album, single, or EP; "Bad Boy" was released on December 10, 1966, on a compilation album, *A Collection of Beatles Oldies*, with 15 other songs, all of which were previously released. This track was available in the U.S. market more than a year earlier, on June 14, 1965, when it was included in *Beatles VI*, and it is shown with the release date for the U.S. market. This track was recorded on the same day as "Dizzy Miss Lizzy," which was subsequently included in the studio album *Help!*. The black points are songs credited to John Lennon–Paul McCartney; red indicates George Harrison as songwriter; the purple originates with other combinations of the Beatles; and the blue points are cover songs.

To prevent songs with exactly the same length and release date from appearing on top of each other, data points are jittered horizontally by a small amount equivalent to four days when necessary. For example, both "And Your Bird Can Sing" and "For No One" from *Revolver* are two minutes and one second long. Also, in two instances, a single was released on the same day as was an album—once with *Beatles for Sale* and another with *Revolver*. These two singles also were moved to the right by four days so the data can be seen.

Observations

The vast majority of songs in the early phase of the Beatles' career (up to *Revolver*) are between two and three minutes in length, perhaps reflecting expectations for radio play at the time. The inclusion of a relatively large number of cover songs during this period also may have been intended to attract a wider audience. All 23 of their cover songs come from this early phase. The end of this phase is marked by their final concert, which took place on August 29, 1966, 24 days after the release of *Revolver*. After this concert, the Beatles started to spend much more time in the studio creating less-traditional and less radio-friendly songs. Songs lasting longer than three minutes became commonplace, although they also released many short songs during this time. Furthermore, Harrison's contribution was much more prominent during this later phase.

Note: The data and R codes are available upon request. Email the author at *tatsuki.koyama@vanderbilt.edu*.



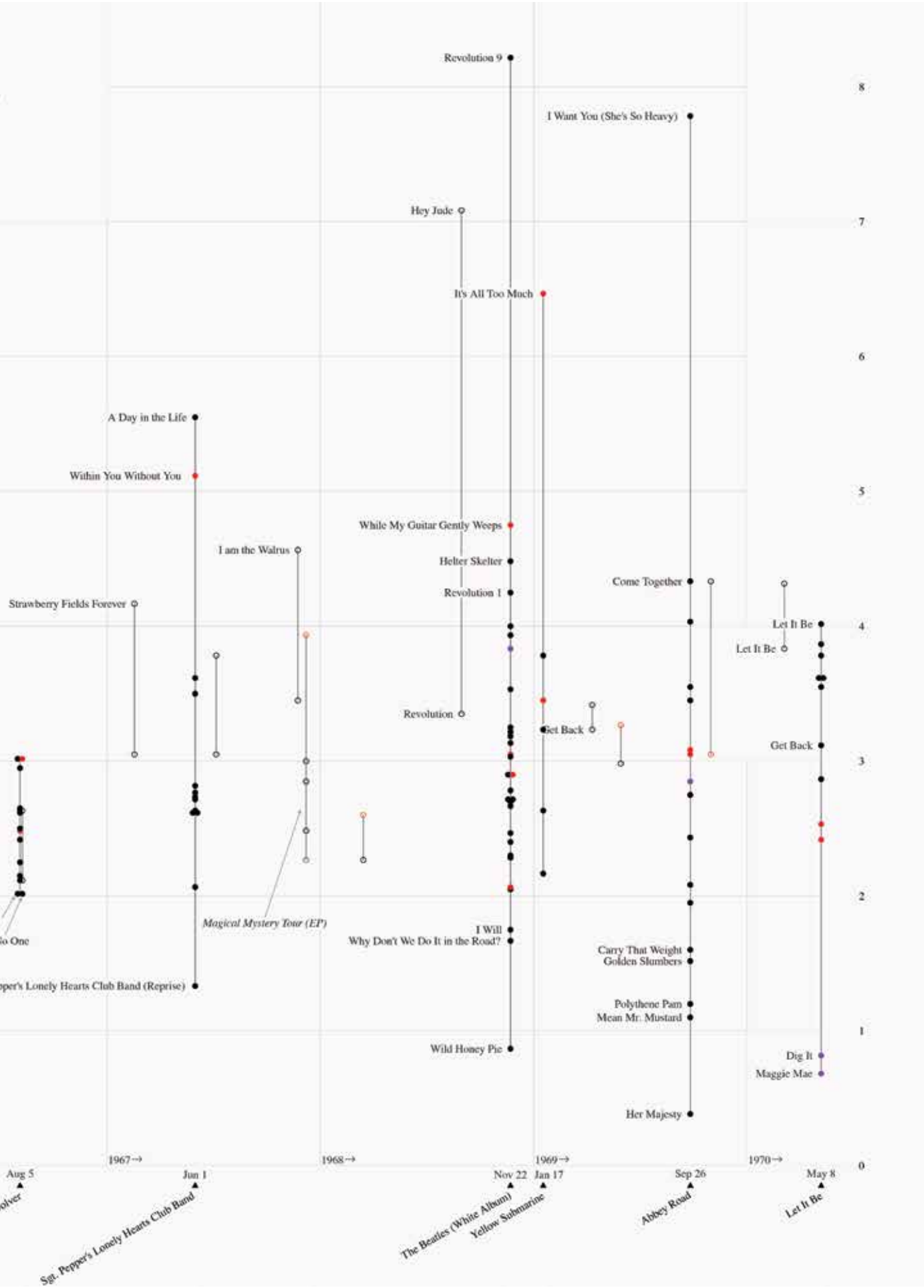

Revolution 9
I Want You (She's So Heavy)
Hey Jude
It's All Too Much
A Day in the Life
Within You Without You
While My Guitar Gently Weeps
I am the Walrus
Helter Skelter
Revolution 1
Come Together
Strawberry Fields Forever
Let It Be
Let It Be
Revolution
Get Back
Get Back
Magical Mystery Tour (EP)
I Will
Why Don't We Do It in the Road?
Carry That Weight
Golden Slumbers
Polythene Pam
Mean Mr. Mustard
Wild Honey Pie
Dig It
Maggie Mae
Her Majesty
8
7
6
5
4
3
2
1
0
1967→
1968→
1969→
1970→
Aug 5
Jun 1
Nov 22
Jan 17
Sep 26
May 8
Sgt. Pepper's Lonely Hearts Club Band
The Beatles (White Album)
Yellow Submarine
Abbey Road
Let It Be

Photo by Grace Forrest, Walk Free Foundation, *www.by-grace.org*.

Unlocking the Statistics of Slavery

Kevin Bales

Statistically, the history of slavery in our world can be divided neatly into two. In the first half, from the very earliest of human writing and records, slaves made up a part of that record. From Sumerian cuneiform to Egyptian hieroglyphs to Greek and Latin scripts and the incorporation of the useful Arabic zero, slavery was a blatant and measurable part of human existence. The clay counting tablets of Mesopotamia, for example, recorded slaves amongst the cattle and grain, and while the papyrus records of the pharaohs rarely survive, the great stone carvings along the Nile enumerate slave captures and clearly assign ownership.

Consider the "Battlefield Palette," (see page 20), thought to originate around 5,200 years ago. Carved into a soft, sedimentary stone, it is considered an important link to the distant past. While most of the story it tells is through pictures, it is thought to be both the earliest depiction of a battle scene and include some of the first representations of the glyphs that, in time, would become the Egyptian writing system of hieroglyphs.

Two of these glyphs are important. One represents the standard or totem that denotes power and authority, and the other is the "man-prisoner," or "captive," glyph. While the meanings of these first written "words" are potent, the picture itself is perfectly clear enough. Note the bound men being marched away at the top left, the hands that control them emerging from the "standard glyph" of power and authority. Below, their slaughtered compatriots have been stripped naked and are being feasted upon by vultures, crows, and a lion. One bound captive has been killed and a bird is pecking out his eyes. Just above them to the right, another captive (seen only from the waist down) is being marched away, his hands tied

This article originally appeared in *CHANCE* 30.3.

Battlefield Palette, thought to originate around 5,200 years ago.

behind his back. Driving him along is the only fully clothed figure in the stone.

When Rome grew into an empire, its economy running on slavery the way the United States today runs on oil, the counting, buying, selling, transferring, giving, and inheritance of slaves must have filled entire record halls. When David Eltis and David Richardson began their project to illustrate the entire trans-Atlantic slave trade over a 366-year period (1501–1867), they found surviving records covering four-fifths of all voyages made—34,934 deliveries in the trade that carried some 12.5 million slaves to the New World.

Possibly the last truly accurate measurement of slavery occurred in 1860, when the United States Census enumerated those held in legal bondage. In that year, the total number of slaves was 3,950,529, accounting for 13% of the U.S. population.

A precise count of slaves was crucial, since the Constitution of the United States calculated how many representatives each state could send to Congress based on population. Although they could not vote and were essentially items of property, each slave was included in the population count as three-fifths of a person, thus greatly aggrandizing the voting power of each slave owner as well as assuring that numerically fewer Southerners could match, through their property, the representatives of the more-populous North.

The second half of the statistical history of slavery begins when slavery becomes illegal. From that time and through the rest of the 20th century, however, there has been no reliable measurement of the extent of slavery in any country, with the possible exception of the records of slave laborers kept by the Nazi regime during the Second World War (Allen, 2004). While legal slavery meant formal records were created, the ongoing criminalization of slavery (even when antislavery laws were rarely enforced) meant there were now no, or very few, hidden accounts of slavery. Notable limited exceptions include files kept by social service agencies on escaped slaves, or the very few legal records linked to the arrest of slaveholders.

At the beginning of the 21st century, just as interest and awareness in modern slavery and its supporting conduit activity of human trafficking was growing rapidly, no reliable information existed on the prevalence of slavery—but it is worth noting that there were widely circulating, but baseless, estimates ranging from no slaves in the world to a few million to a much-quoted total of 100 million.

Within this context, I carried out a systematic collection of information from 1994 until 1999 to construct a global estimate (Bales, 1999), with a revised version (Bales, 2002) and a detailed explanation of my methodology (Bales, 2004). This estimate put the number of slaves in the global population at 27 million. This was an estimate, as I made clear, built from secondary source information, processed by a team that assessed each source for validity as far as possible, with care taken to treat each source with skepticism and to record only the conservative end of any range of estimates.

From the beginning, I was highly critical of my data and very much aware of the shortcomings of the data. I noted, for example, that I was "potentially building upon bad estimates to construct worse ordinal or ranking estimates. Even worse…there was no way to know if this was the case or not." That being the situation, and with that and other provisos, I made my data freely available, leading to its use by the statistician and methodologist Robert Smith and others.

Also at the beginning of the 21st century, a number of other groups and individual scholars began attempting to measure slavery in local areas, nations, regions, and globally. In doing so, they quickly divided into two groups and then, in parallel, proceeded through four stages of methodological development.

These two groups were divided by their approach to data transparency, reproducibility, and replication. Some researchers, primarily social scientists in academic appointments and some nongovernmental organizations (NGOs), operated on the basis that it was important to make their data freely available in ways that would allow other researchers to test, replicate, and potentially reproduce their results in commensurate studies—thus adhering to one of the fundamental principles of the scientific method.

The other group, for a number of reasons, did not feel able to share their data freely. This was sometimes due to political sensitivities, or notions or requirements of proprietary interest in the data collection, concerns about the data itself, or the methods of collection or analysis. Government sources, in particular, were loath to make data public. So were commercial organizations whose business model was to use the freely available data to construct indices and synthetic reports that they then sold to clients, but which were not transparent about data origins.

It is worth noting that transparency, replication, and reproducibility are issues of increasing concern more broadly across the sciences. A recent article in *Nature* and a recent report on biomedical research both point to a growing unease over the lack of data sharing and replication.

While there is general disquiet over this issue, there is clear consensus about its remedies: 1) openly sharing results and the underlying data with other scientists; 2) collaboration with other research groups, both formally and informally; 3) publicly publishing the detail of study protocols; and 4) reporting guidelines and checklists that help researchers meet certain criteria when publishing studies. At the time of writing this article, no consensus has arisen concerning the practice of data transparency in the field of the measurement of slavery prevalence, nor have reporting guidelines been agreed upon and set for slavery researchers.

Within this context, the first stage in the measurement of contemporary slavery, exemplified by my work, relied upon secondary sources, including governmental records, NGO and service provider tallies, and reports in the media; in short, any source that might shed light on the extent of slavery. Even when sources were systematically assessed for reliability, these estimates (Bales, 2004; ILO, 2005) could only be seen, at best, as an approximation of the global situation.

One expansion of this method (*Hidden Slaves*) in the United States was an attempt to triangulate secondary sources with surveys of service providers and government and law enforcement records. While the estimates derived in this first methodological stage were not widely different from each other, it was impossible to ensure their comparability or validity.

The second stage was set in motion by the pioneering work of Pennington, Ball, Hampton, and Soulakova in 2009. This team introduced a series of questions concerning human trafficking into a random sample health survey of five Eastern European countries (Belarus, Ukraine, Moldova, Romania, and Bulgaria). Employing random sample surveys, they were able to build the first representative estimate of the proportion of each country's population that had been caught up in human trafficking.

It is worth noting that the terms "modern slavery" and "human trafficking" are sometimes used interchangeably, trafficking is simply one of many processes by which a person might be brought into a state of enslavement. While the human trafficking process suffers from being defined in different ways in a number of legal instruments and operational definitions, it is normally understood to mean the recruitment and then movement of a person into a situation of enslavement and exploitation.)

The work of Pennington, et al., was critical to advancing the measurement of the prevalence of slavery for two reasons. Firstly, it demonstrated that, at least in some countries and circumstances, enslavement could be measured through random sample surveys of the full population. Secondly, by fixing valid data points for these five countries, it became possible to begin the process of estimating the *range* of modern slavery in countries by using these, and other emerging survey results, to extrapolate the prevalence of slavery in other countries (Datta and Bales).

In addressing the question of range, it had become clear by 2009 that cases of slavery (although not measures of slavery prevalence) were being reported in virtually all countries with populations over 100,000 (Bales, 2004; UN-GIFT, 2009). For that reason, the low end of the global range of prevalence, for countries in which measurement was possible, could be assumed to be greater than zero.

In the same year, a U.S. Agency for International Development (US-AID) and Pan American Development Foundation report included a random sample survey

of child *restavek* trafficking and slavery in Haiti. This form of the enslavement of children into domestic service and other types of exploitation had been widely investigated (Cadet and Skinner), in part because of its ubiquity in urban settings, but never estimated through surveys.

This US-AID survey estimated that 225,000 children were enslaved in Haitian cities, equaling 2.3% of the national population. This estimated proportion of the Haitian population was assumed to be in the upper range of the global distribution of slavery prevalence for two reasons. The first was that most investigators had noted the pervasive nature of this form of slavery as compared to slavery in other countries; the second was that of the few existing representative sample measures of slavery by country, Haitian *restavek* slavery was, by far, the largest.

The culmination of this second stage came with an emerging sense of the range of prevalence across countries and an increase in the amount of data available from random sample surveys. In addition to data from the Pennington, et al., and Haiti surveys, random sample surveys of slavery were also identified in three more countries (Niger, Namibia, and the Democratic Republic of Congo; "Namibia Child Activities Survey" and Johnson, et al., 2010). The combination of these disparate surveys, and their use in building an extrapolation estimation process, generated the 2013 global estimate of 29.8 million slaves in the first edition of the Global Slavery Index.

The third stage in the estimation of the prevalence of slavery came with the introduction of systematic and *comparable* representative random samples in a number of countries. In late 2013, the Walk Free Foundation commissioned seven national surveys (Pakistan, Indonesia, Brazil, Nigeria, Ethiopia, Nepal, and Russia), using the Gallup International World Poll. These comparable surveys were rolled into the iterative extrapolation process that generated a global estimate of 35.8 million people in slavery worldwide, published in the 2014 Global Slavery Index.

The World Poll survey data are representative of 95 percent of the world's adult population. The World Poll uses face-to-face or telephone surveys, conducted through households (where a household is defined as any abode with its own cooking facilities, which could be anything from a standing stove in the kitchen to a small fire in the courtyard) in more than 160 countries and more than 140 languages. The target sample is the entire civilian, non-institutionalized population, aged 15 and older.

With the exception of areas that are scarcely populated or present a threat to the safety of interviewers, samples are probability-based and nationally representative. The questionnaire is translated into the major languages of each country, and field staff conduct in-depth training and receive a standardized training manual. Quality control procedures ensure that correct samples are selected and the correct person is randomly selected in each household. A detailed description of the World Poll methodology is available online: *http://bit.ly/2u9v6Iz*.

This mixture of comparable representative surveys using the same format and wording, and the "found" surveys, each unique in design and sampling, were used to build an extrapolation process that also included a series of variables measuring a range of factors that might predict vulnerability or propensity to slavery within a country. In many ways, this introduction of an extrapolation process for estimating slavery in those countries without direct surveys, linked to a number of predictors of enslavement, was the platform for building the fourth stage of prevalence estimation.

It is important to contextualize this fourth stage, since all previous stages lead to this system of longitudinal and iterative testing of prevalence measures. Slavery estimation had moved from secondary source "guesstimation" to comparable random sample surveys to an algorithmic process ensuring comparability and the potential for replication, reproducibility, and further "ground-truthing" research. Because this fourth stage of testing can continue to be elaborated over much iteration, it is unlikely a fifth stage will emerge in the near future.

Given, as well, that this technique can also be combined with Multiple Systems Estimation (MSE) (van Dijk and van der Heijden provide a full discussion of MSE in this issue) to generate prevalence measures for highly developed nations for which surveys are not appropriate, it is possible to imagine a global estimate in which most country estimates rest on a firm quantitative methodological foundation.

If there is a fly in this optimistic ointment, it is that while measurement issues are slowly being resolved, little progress has been made in arriving at a shared *operational definition* for the object of study: slavery.

The Definitional Challenge

It is worth noting that, for most of human history, slavery was both ubiquitous and undefined; slavery was so common that defining it was not necessary. Over time, laws did set out who might be enslaved or manumitted, but it was an activity so well understood that

it was rarely given a precise definition. In a number of historical contexts, detailed criteria set out who might be enslaved, such as the Slave Codes of the U.S. Deep South, Roman slave laws, or even Nazi Nuremberg—Reich Citizenship—laws, which allowed the separation within the population of people without rights.

These, however, are not definitions in a human rights framework, but tools designed by slaveholders to specify and control the enslaved and/or enslavable. Nor was slavery normally defined in the early treaties and laws that regulated and then abolished legal slavery in the 18th and 19th centuries.

For example, the 13th Amendment to the U.S. Constitution simply reads, "Neither slavery nor involuntary servitude…shall exist within the United States." It was only in the 20th century, when virtually all slavery was ostensibly illegal, that it was felt that the human activity known as "slavery" needed a specific definition. This perceived need was exacerbated in the early 1990s, when a mushrooming of human trafficking paralleled an equally growing traffic in arms and drugs across the same borders.

In response to this suddenly visible movement of trafficked people into developed countries, especially into commercial sexual exploitation, a number of groups and political actors pressed for new regulation. Some commentators describe a "moral panic" in this period, pushed by diverse groups. A key outcome of this sudden interest and energy was a number of new international conventions and national laws, all of which tended to define slavery or human trafficking differently.

This is not the place to review all these varying definitions, but it is worth noting, as an example of the mix of definitional frameworks, that some activities, such as forced or compelled marriage or organ trafficking, were defined as subsets within slavery and others were not. Still other new legal definitions defined slavery itself as a subset of another activity, such as human trafficking. The lack of agreement between these legal instruments has created confusion across jurisdictions and generated a lack of conceptual clarity when confronting activities that may or may not be considered within the wider category of slavery.

A second result is that courts have issued rulings that either set down divergent definitions or interpret the same definition very differently. Remarkably, international law also says that the prohibition of slavery is *jus cogens*—an internationally applicable peremptory norm from which no derogation is ever permitted. Thus, we find ourselves with a universally and comprehensively forbidden crime, but one that is defined in different, often even contradictory, ways.

These disparities in legal definitions create difficulties in developing an operational definition for another reason: The voices and views of those who have been enslaved have been excluded in their construction. After all, slavery is, first and foremost, a lived experience—not a legal definition, an analytical framework, or a philosophical construct. At the moment it is occurring, slavery is first the experience of an individual person, and second a relationship between at least two people: the slave and the slaveholder. Slavery also carries cultural, political, and social meanings; meanings that are important to understand if we are to grasp the context of slavery and the factors that might predict its occurrence.

Within these different dimensions of enslavement, the lived experience of slaves is of primary importance. not least because the way in which slavery is classified and defined, in law and in public opinion, determines who is eligible for relief and who is not; who may live with some measure of personal autonomy and who may die in bondage.

It has been necessary to discuss the legal definitions of slavery as a preamble to understanding the lack of a generally accepted operational definition of slavery because of controversy, and misunderstanding, within the larger anti-slavery field concerning how slavery is defined. Many actors in the larger academic, as well as the applied anti-slavery movement, have argued that, because slavery is now an illegal activity, legal definitions must be paramount. But legal definitions are written for a specific purpose: to guide the implementation of law; to make clear, within the legal framework, when a specific crime has been committed, and that is not, the aim of an *operational definition*.

An operational definition aims to identify, in a precise way, the nature and characteristics of an object of scientific research. It is fundamentally a definition that sets out clearly what is, and what is not, the subject of inquiry and measurement. Attempts to use any of the widely disparate legal definitions to guide research into the social activity known as slavery have not been illuminating or successful—with one exception.

Over a three-year period (2010–2012), a group of legal, social science, and other experts met to resolve the definitional confusion, asking if there were a definition that might both apply and be useful within the law and operationally to guide social science, and especially quantitative, research. The resulting consensus of this group was that the definition

of slavery available in the existing international legal framework that provided the clarity and usefulness was that in the 1926 Slavery Convention of the League of Nations: "Slavery is the status or condition of a person over whom any or all of the powers attaining to the right of ownership are exercised."

The committee of experts added explanatory guidelines to clarify the application of the 1926 definition. The aim was to elucidate the "powers attaching to the right of ownership" so the attributes of any instance of suspected enslavement might be compared to the criteria inherent to the 1926 Convention. To accomplish that, it is necessary to, firstly, locate the legal definition in the lived reality of enslavement, and, secondly, specify the attributes of ownership more clearly that apply to the law of property and make clear how these attributes apply to the situation of enslavement.

The core of this adaptable definition is *the powers attaching to ownership*. The most-central of these is *the right to possess*—according to Honoré, this is "the foundation on which the whole superstructure of ownership rests" (1961). Possession is demonstrated by control—normally, exclusive control. This is best demonstrated in what Hickey describes as the "maintenance of effective control" (2010), meaning exercising control over time and likely to include other instances or indicators of ownership.

These other instances are *the right to use*, *right to manage*, *right to income*—"use" being the right to enjoy the benefit of the possession; "manage," the right to make decisions about how a possession is used; and "income," the right to profits generated by a possession. In addition to these central rights of ownership is *the right to capital*, which refers to the right to dispose of the possession by transfer, consumption, or destruction.

These "instances of ownership"—control, use, management, and profit—may be regarded as the central rights of ownership. It is their presence and exercise that can be applied and tested within a situation, such as slavery, where actual legal possession is not permitted. Given the illegality of slavery in all countries, they provide a critical power of definition and identification of the crime of slavery—these "instances" can be treated as measurable indicators in an operational definition of slavery.

The other attributes of possession, as normally expressed, pertain primarily to ownership that is sanctioned by law and so are less useful in understanding modern forms of illegal enslavement. That is not to say, however, that modern slaveholders do not seek to exercise these "rights" when they can. These other attributes of possession include *rights of security*—protection against illegal appropriation of a possession; *transmissibility*—the right to transfer legal ownership; and two indicators of the permanence of possession: *absence of term*—the lack of a time restriction on ownership (an attribute in that slavery is a relationship of control that exists for an indeterminate period of time); and the *residual character* of ownership—a possession may be loaned or rented, but will return to its owner and never cease to be property.

The key product of the committee of experts was the *Bellagio-Harvard Guidelines on the Legal Parameters of Slavery*. This short document sets out how the definition of slavery used in the 1926 Convention is coherent and useful in both legal and social science contexts.

It is unlikely that there will be a consensus on a shared operational definition among researchers into slavery any time soon, but at the very least, a discussion and exploration of such potential operational definitions should occur. The arguments offered for the "exceptionalism" of certain types or methods of slavery tend to create studies and reports that are non-comparable; that is, they "compare apples and oranges." The fundamental result is a growing body of literature that is much less useful in addressing the crime of slavery than it might be.

If the definitional problem were not sufficiently challenging, it is exacerbated by the lack of transparency and reproducibility in research methods and data.

The Need for Transparency and Reproducibility

If the social sciences have achieved a basic set of methodologial tools with which to measure slavery, and an operational definition that might guide and achieve comparable research on contemporary slavery, a serious challenge remains in a lack of data transparency, which makes the fundamental scientific requirement of reproducibility impossible. In many ways, it is surprising that such a lack of transparency exists, given the nature of the phenomenon being studied.

Slavery and human trafficking are serious crimes, with terrible repercussions on the lives of the enslaved. The deaths, diseases, injuries, and mental health impacts of slavery are well-known. Slavery is a threat to life, health, well-being, and the social stability of communities, as well as a known facilitator of conflict, rape, violence in many forms, and brutal treatment of children. Both the

immediate effects of slavery and its *sequelae* across not only generations but time are also well known.

Given those demonstrated and widely known facts, the ongoing lack of transparency, and the lack of data sharing in the study of slavery, is not just a threat to good science. It prevents comparable analyses that might reduce suffering and the extreme human cost of slavery.

Shared data have the potential of leading to the amelioration and reduction of this horrific crime. For that reason, a quick review of the nature of why science operates through open dialogue and the sharing of data and results, and why that practice is critically needed today in the study of slavery, is necessary.

Within the sciences, including the social sciences, the internal political economy—the measurement of worth and meaning—is not financial. It is much closer to what anthropologists call a "gift economy." Spufford provided a wonderful explanation of the gift economy in the medical sciences: "In a gift economy, status is not determined by what you have, but by what you give away. The more generous you are, the more you are respected; and in turn, your generosity lays an obligation on other people to behave generously themselves, to try to match your generosity and so claim equal or greater status.…When scientists practice [their gift economy], the gift they give away is information" (2003).

While there are informal expectations within the academic gift economy, it is also rigorously and formally governed by the rules of scientific publishing. These rules include requirements that published articles must make data freely available for re-analysis, and that sources of data and ideas are clearly acknowledged and cited.

It is important to note that these rules do not hamper competition; in fact, they increase it and foster it, since giving everyone access to the same shared information and data doesn't just level the playing field; it opens the field to any and all comers. This competition can be harsh, energetic, even bruising, but that is also a reflection of the fact that the reward for competing successfully is nothing as mundane as money. It is a much more powerful motivator: respect.

Of course, if the only reason for transparency and reproducibility was to gain respect in a circular game of of academic one-upmanship, there would be little point to observing such rules, but the highly productive scientific gift economy is only a foundation for a much-more-important and pragmatic activity.

Science is based upon the accretion of ideas and findings. Every scholar may believe their ideas and findings are important, but more widely, in society as a whole, certain ideas and findings are considered critically important and valuable in their power to transform or protect human life. Medical research is a clear example, and the hoarding of a new idea or data with the potential to save lives or reduce suffering would be seen as not just unacceptable, but shameful.

So, too, it must be argued, would be withholding ideas, findings, or data pertaining to a locus of suffering, a crime as monstrous as slavery. When businesses seek to monetize information about slavery as a condition of their business plans, they have to lock away ideas and data, since free data cannot be monetized. When non-governmental organizations seek to lock away and control data, for whatever reason, they place themselves in the same category of selfish negligence as such businesses, since ideas and data withheld cannot be used to solve pressing problems, reduce suffering or, even, free slaves.

In many areas of research with a direct impact on lives and well-being, shared systems for information and data exchange are common. The open and freely searchable European Bioinformatics Database, for example, hosts a whole series of separate specialist databases. One of these alone, the Malaria Data site, holds records of 371,255 compounds and 25,726 publications.

The systematic study of contemporary slavery is relatively recent, but the destructive potential of the object of study suggests that a system of information and data exchange is overdue. In the same way that the scientific study of slavery is hampered by definitional confusion, it is also held back by a failure to respect the rules of science. In some arcane areas of academic endeavor, that might not matter, but slavery—for obvious reasons—is not one of them.

This is why [the special] issue of *CHANCE* [was] not simply useful to scholars, but important in the wider sense. The articles in [that] issue seek to achieve two key goals. The first is to make clear the current state of play in the field of measuring slavery; the second is to demonstrate what can be achieved when researchers in this field operate by the shared rules of scientific endeavour. All of the authors are keenly aware that they are working in a new field; that they are, at times, setting out new ideas, procedures, and methods, and most of all, that to make progress, they must do so in a way that is transparent and open to critique and improvement.

The work presented in this special edition on developments in vulnerability modeling and how that might be used in an extra-polation

process to estimate the prevalence of slavery is both groundbreaking and a work in progress. The explanation of the use of methodologically sophisticated Gallup World Poll surveys to measure slavery prevalence better explores what happens when a trusted and refined tool is brought to bear on a hidden human activity.

Since much of the information available globally is the product of governments, it is crucial to assess the reliable of such data, and what tools might be used to resolve data integrity. The exploration of the innovative application of the technique of MSE to measuring the prevalence of slavery appears to offer a solution to the problem of estimating the extent of slavery in well-developed countries.

The final article [in the special issue] suggests not just the way forward, but the tools and practices that will be needed to move forward expeditiously; ideally into a world where the metrics of slavery are used to guide the eradication of slavery. C

Further Reading

Academy of Medical Sciences. 2015. Reproducibility and reliability of biomedical research: improving research practice, Symposium Accessed October 20, 2016, at *http://bit.ly/2vYDgZS*.

Allen, M.T. 2004. *Hitler's Slave Lords: The Business of Forced Labour in Occupied Europe*, Chapel Hill: University of North Carolina, The History Press.

Baker, M. 2016. Is there a reproducibility crisis? *Nature* 535:452.

Bales, K. 1999. *Disposable People: New Slavery in the Global Economy*. University of California Press.

Bales, K. 2002. The Social Psychology of Modern Slavery. *Scientific American*.

Bales, K. 2004. International Labor Standards: Quality of Information and Measures of Progress in Combating Forced Labor. *Comparative Labor Law and Policy* 24(2).

Cadet, J-R. 1998. *Restavec: From Haitian Slave Child to Middle-Class American*. Austin: University of Texas Press.

Datta, M.N., and Bales, K. 2013. Slavery in Europe: Part 1, Estimating the Dark Figure, *Human Rights Quarterly* 35.3.

Eltis, D., and Richardson, D. 2010. *Atlas of the Transatlantic Slave Trade*. New Haven: Yale University Press.

Hickey, Robin. 2010. Exploring the Conceptual Structure of the Definition. Paper presented at Slavery as the Powers Attaching to the Right of Ownership, Bellagio, Italy.

Hidden Slaves: Forced Labor in the United States. 2004. National report to the International Labor Office for the Global Survey of Forced Labor, with the Human Rights Center. Berkeley: University of California Berkeley.

Honoré, A.M. 1961. Ownership, in A.G. Guest (ed.). *Oxford Essays in Jurisprudence*. Oxford, England: Oxford University Press.

International Labor Office. 2005. Report of the Director General: *A global alliance against forced labor*, Global Report under the Follow-up to the ILO Declaration on Fundamental Principles and Rights at Work, Geneva, Switzerland.

International Labor Organization. *ILO Global Estimate of Forced Labor in 2012: Results and Methodology*. Geneva, Switzerland: ILO.

Johnson, K., Scott, J., Rughita, B., Kisielewski, M., Asher, J., and Ong, R. 2010. Association of Sexual Violence and Human Rights Violations with Physical and Mental Health in Territories of the Eastern Democratic Republic of the Congo. *Journal of the American Medical Association* 304(5).

Ministry of Labour and Social Welfare, Directorate of Labour and Market Services, 2009. Namibia Child Activities Survey (2005), ISBN: 978-086976-787-0; *Enquete Nationale sur le Travail des Enfants au Niger.* (ILO), accessed 23/09/14.

Pierre, Y-F., Smucker, G.R., and Tardieu, J-F. 2009. Lost childhoods in Haiti: Quantifying child trafficking, *restavèks*, and victims of violence. International Development and Pan American Development Report. *http://bit.ly/1N6TLVv*.

Pennington, J.R., Ball, W.A., Hampton, R.D., and Soulakova, J.N. 2009. The Cross-National Market in Human Beings, *Journal of Macromarketing* 29(2):119–134.

Skinner, E.B. 2008. *A Crime So Monstrous*, New York: Mainstream Publishing.

Smith, R.B. 2009. Global human development: accounting for its regional disparities. *Quality and Quantity* 43(1).

Spufford, F. 2003. *The Backroom Boys: The Secret Return of the British Boffin*. London: Faber & Faber.

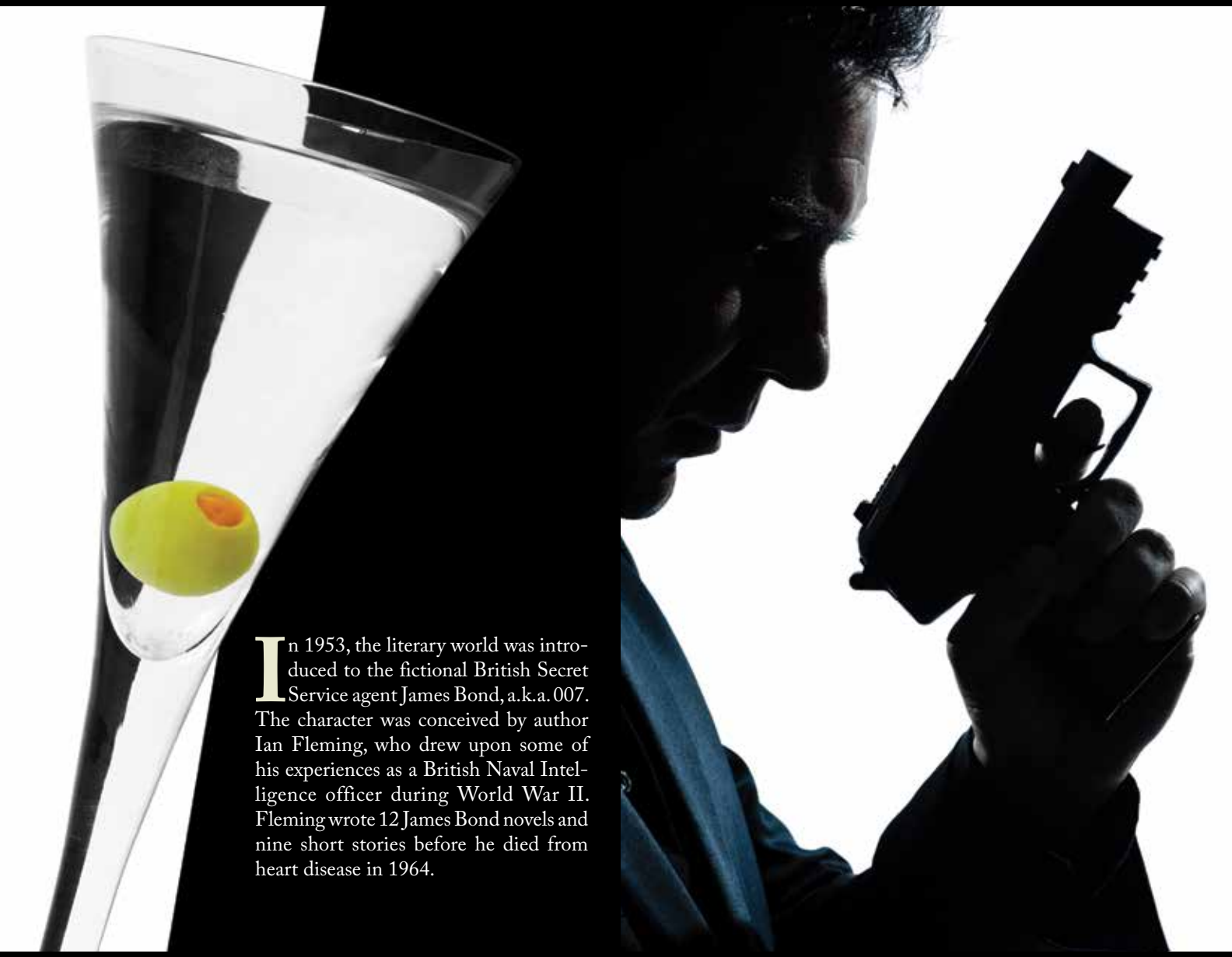
United Nations Global Initiative to Fight Human Trafficking. 2009. *Global Report on Trafficking in Persons*. Vienna, Austria: U.N. Office on Drugs and Crime.

About the Author

Kevin Bales, PhD, CMG, FRSA, is the Professor of Contemporary Slavery at the University of Nottingham, research director of the Rights Lab, and lead author of the Global Slavery Index. He was co-founder of the NGO Free the Slaves. His book *Disposable People: New Slavery in the Global Economy* was published in 11 languages and named one of "100 World-Changing Discoveries" by the Association of British Universities. The film version won a Peabody and two Emmy awards. His newest book, *Blood and Earth: Modern Slavery, Ecocide, and the Secret to Saving the World* (2016), proves the deep link between modern slavery and climate change.

Bond. James Bond.
A Statistical Look at Cinema's Most Famous Spy

Derek S. Young



In 1953, the literary world was introduced to the fictional British Secret Service agent James Bond, a.k.a. 007. The character was conceived by author Ian Fleming, who drew upon some of his experiences as a British Naval Intelligence officer during World War II. Fleming wrote 12 James Bond novels and nine short stories before he died from heart disease in 1964.

This article originally appeared in *CHANCE* 27.2.

Fleming's 007 novels achieved moderate success and critical acclaim; however, what catapulted James Bond into the pantheon of cultural icons was the introduction of the character portrayed by Sean Connery in 1962's "Dr. No." As of 2013, there have been 23 official James Bond films produced and released by Eon Productions, six actors have donned the tuxedo as 007, and numerous cultural influences can be attributed to the secret agent. The films have enjoyed enormous success and popularity spanning their 50-year tenure. When not adjusting for inflation, the Bond franchise is ranked number two on the list of most successful film franchises worldwide, behind the "Harry Potter" films and above the "Star Wars" films. In fact, it has been estimated that around 20% of the world's population has seen at least one James Bond film.

James Bond also has been the subject for academic discourse. For example, Christoph Lindner edited the 2010 book *The James Bond Phenomenon: A Critical Reader*, which is an essay collection providing theoretical perspectives on James Bond as a cultural figure. Daniel Savoye's 2013 book, *The Signs of James Bond: Semiotic Explorations in the World of 007*, provides a critical analysis of the formulaic elements that comprise James Bond films. There also have been university topics courses that explore themes and issues conveyed through the James Bond films and novels. Such courses have been offered at Missouri State University, Thiel College, and the University of Michigan.

In this article, we analyze the James Bond films from a statistical perspective and provide data visualizations of various aspects of the films. We build a multivariate regression model to dissect the box office totals (adjusting for inflation) and average ratings when considering other variables about the films. Using our model, we provide a prediction region for the box office gross and average rating score of the next James Bond film. We also provide a ranking of the actors who have played James Bond in an attempt to answer an often-debated question that even Q has surely struggled with: Who is the best Bond?

The Data

As noted earlier, we only include the "official" movies produced by Eon Productions. In addition to these 23 movies, there have been a few "unofficial" James Bond movies made. The unofficial films are the 1954 made-for-TV movie "Casino Royale," the 1967 comedic spoof "Casino Royale" starring David Niven, and the 1983 competing film "Never Say Never Again," which is a remake of "Thunderball" and stars Connery.

Table 1 shows some of the data for the movies we use in our analysis. The table includes the year each film was released, the name of the film, which actor portrayed James Bond, worldwide gross for the film (in $1,000s) adjusted for inflation, the budget for the film (in $1,000s) adjusted for inflation, and the average online ratings for the films. The original box office totals and budgets were retrieved from *www.the-numbers.com*, a website that maintains box office history and other statistics about movies. These values were adjusted to 2013 dollars using the U.S. Bureau of Labor Statistics' inflation calculator, which uses the Consumer Price Index and was accessed on April 3, 2013. The average user ratings were pulled from each movie's webpage on *www.imdb.com* and *www.rottentomatoes.com*, both of which were also accessed on April 3, 2013. These websites allow users to rate movies on a scale from 1 to 10. For the James Bond films, the user ratings between these two websites are highly correlated, with a Pearson correlation of +0.8868. Thus, for our analysis, we average the pairs of user ratings.

In addition to the variables presented in Table 1, we consider the following about each film for our analysis:

- Length of the theatrical release
- Number of "conquests" James Bond had
- Number of martinis Bond had
- Number of times "Bond. James Bond" was said
- Number of kills by Bond
- Number of kills by characters other than Bond
- Whether the theme song peaked within the top 100 on the UK Singles Chart and the U.S. Billboard Hot 100

The values for the number of kills by Bond and the other characters were obtained from data published on a data blog at the *Guardian's* website (*www.guardian.co.uk*).

Table 1—Official James Bond Films by Release Year

Year Released	Movie	Bond Actor	Adjusted Worldwide Gross (in $1000)	Adjusted Budget (in $1000)	Average Rating
1962	Dr. No	Sean Connery	457,928	7,688	7.50
1963	From Russia with Love	Sean Connery	598,624	15,174	7.75
1964	Goldfinger	Sean Connery	935,404	22,468	8.10
1965	Thunderball	Sean Connery	1,040,693	66,333	6.90
1967	You Only Live Twice	Sean Connery	775,740	66,035	6.60
1969	On Her Majesty's Secret Service	George Lazenby	518,736	50,608	6.75
1971	Diamonds are Forever	Sean Connery	664,969	41,274	6.50
1973	Live and Let Die	Roger Moore	846,046	36,603	6.35
1974	The Man with the Golden Gun	Roger Moore	459,623	32,965	5.90
1977	The Spy Who Loved Me	Roger Moore	710,290	53,636	6.95
1979	Moonraker	Roger Moore	672,514	99,134	5.95
1981	For Your Eyes Only	Roger Moore	498,812	71,514	6.55
1983	Octopussy	Roger Moore	437,059	64,102	5.90
1985	A View to a Kill	Roger Moore	329,322	64,730	5.45
1987	The Living Daylights	Timothy Dalton	390,758	81,749	6.50
1989	Licence to Kill	Timothy Dalton	292,392	78,637	6.25
1995	Goldeneye	Pierce Brosnan	542,985	91,404	7.05
1997	Tomorrow Never Dies	Pierce Brosnan	491,098	159,117	6.20
1999	The World Is Not Enough	Pierce Brosnan	504,091	188,130	6.00
2002	Die Another Day	Pierce Brosnan	557,433	183,255	6.05
2006	Casino Royale	Daniel Craig	686,784	117,465	7.85
2008	Quantum of Solace	Daniel Craig	638,035	248,014	6.40
2012	Skyfall	Daniel Craig	1,120,980	202,240	8.00

Note: The worldwide gross and budget are both adjusted to 2013 dollars.

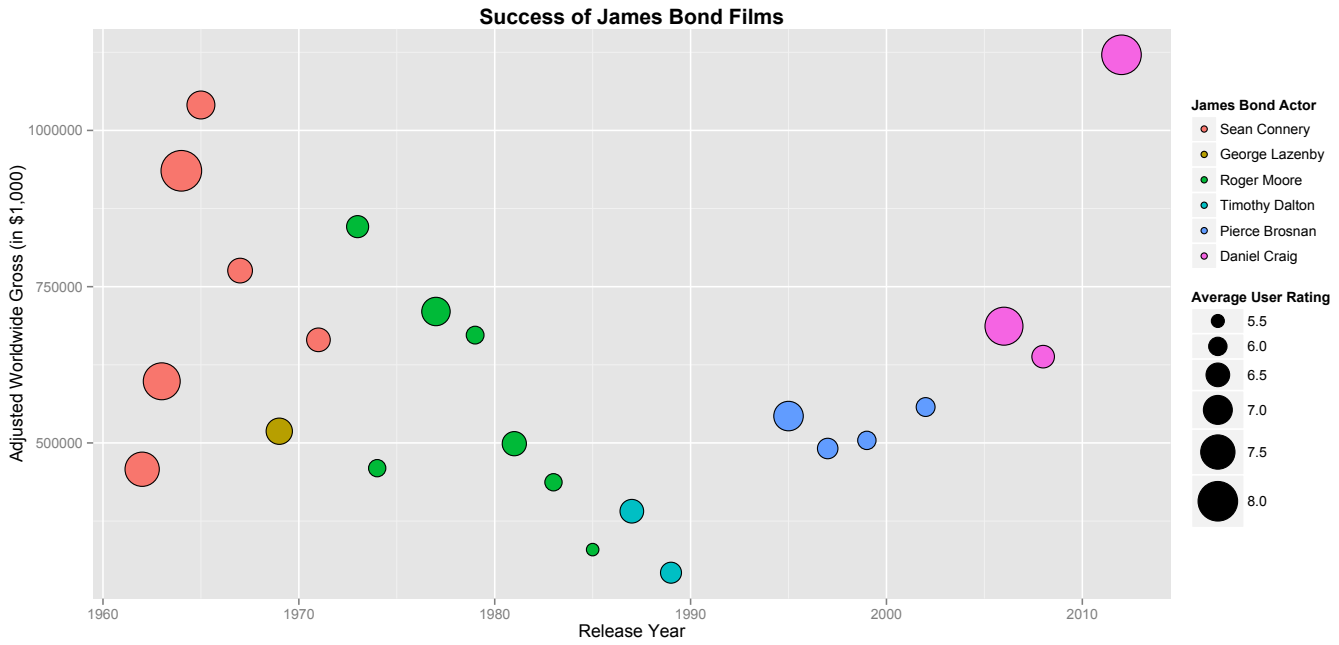


Figure 1. Bubble plot showing the adjusted worldwide gross of James Bond films by the year they were released.

Exploratory Data Analysis

A timeline of the adjusted worldwide gross is given as a bubble plot in Figure 1, where the scale of the bubbles reflects the average user rating of the film. From this plot, we note a few interesting aspects about the James Bond films. First is that with the exception of the lone George Lazenby film, the first film by each actor had higher adjusted worldwide gross sales than the last film by the previous actor. We can see that Roger Moore, whose tenure as James Bond was the longest with seven films, tended to decrease across time in both adjusted worldwide gross sales and average user ratings. The Pierce Brosnan films tended to be consistent in terms of the adjusted worldwide gross sales. Finally, we see that the most recent Daniel Craig film, "Skyfall," is currently the highest-grossing James Bond film and has the second-highest average user rating, just behind Connery's "Goldfinger."

We also provide a stacked area chart of the cumulative sales of the James Bond films in Figure 2. This chart is in terms of millions of U.S. dollars, with the sales are adjusted to 2013 dollars. We see that the cumulative U.S. sales have increased consistently over time. However, the total world sales have increased at a greater rate relative to just the U.S. sales, especially given the success of "Skyfall." For example, as of 2013, "Skyfall" was the highest-grossing film of all time in the United Kingdom.

Next, we look at two vices associated with James Bond: martinis and conquests of female companions. Figure 3 is a bar chart showing the distribution of these two vices by each James Bond actor. Overall, the average number of conquests per Bond is similar among the actors. Moore has the highest number, but he also starred in the most films. The distribution of martinis by each actor is noticeably more variable. For example, Moore has the lowest average number of martinis while Craig has both the highest total and average number of martinis. In fact, Craig's James Bond is the only one whose martini count exceeds his conquest count.

While James Bond is infamous for his vices, he is also known for his license to kill. Figure 4 is a bubble plot showing each James Bond actor's kill ratio (i.e., the average number of kills per film) versus his total kills in all films. The bubbles are proportional to the kill ratio by all other characters in the corresponding films. For example, we see the kill ratio by other characters always exceeds that of the James Bond actor, except Brosnan. The biggest differential between kill ratios is with Connery, while clearly Brosnan had the highest kill ratio and total kills. Thus, Brosnan has the dubious distinction of being the deadliest Bond.

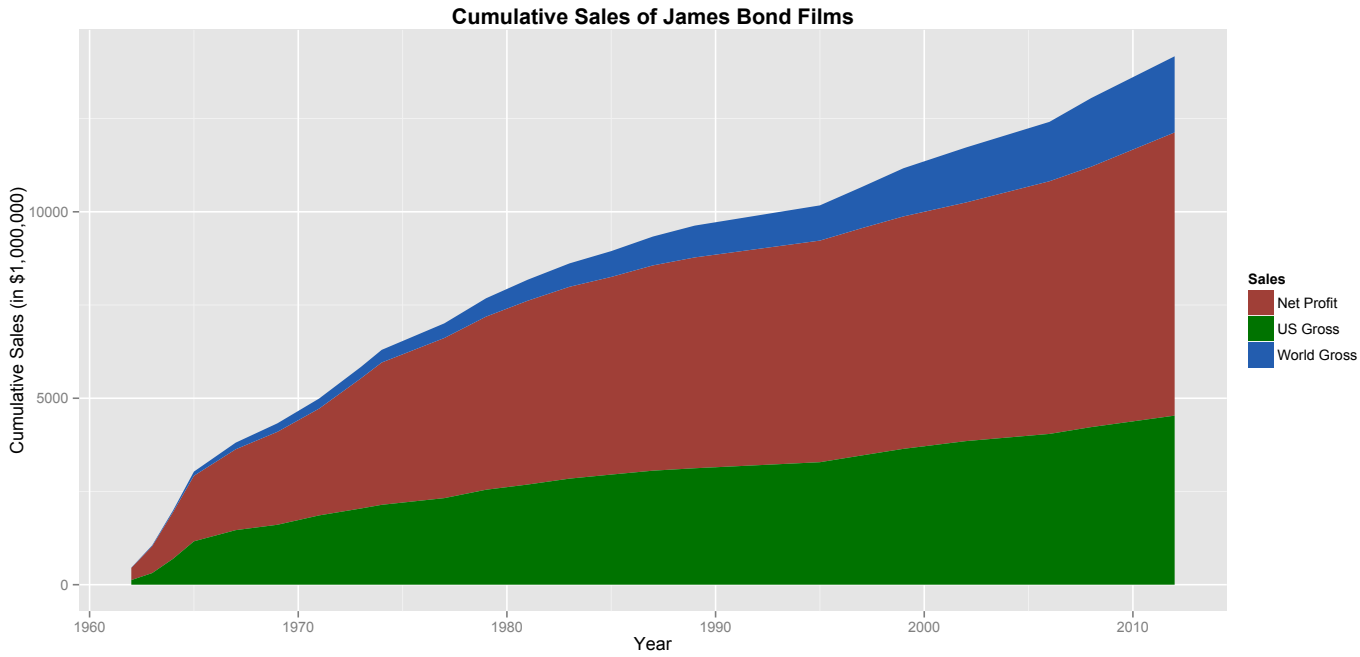


Figure 2. Stacked area chart of cumulative sales of the James Bond films.

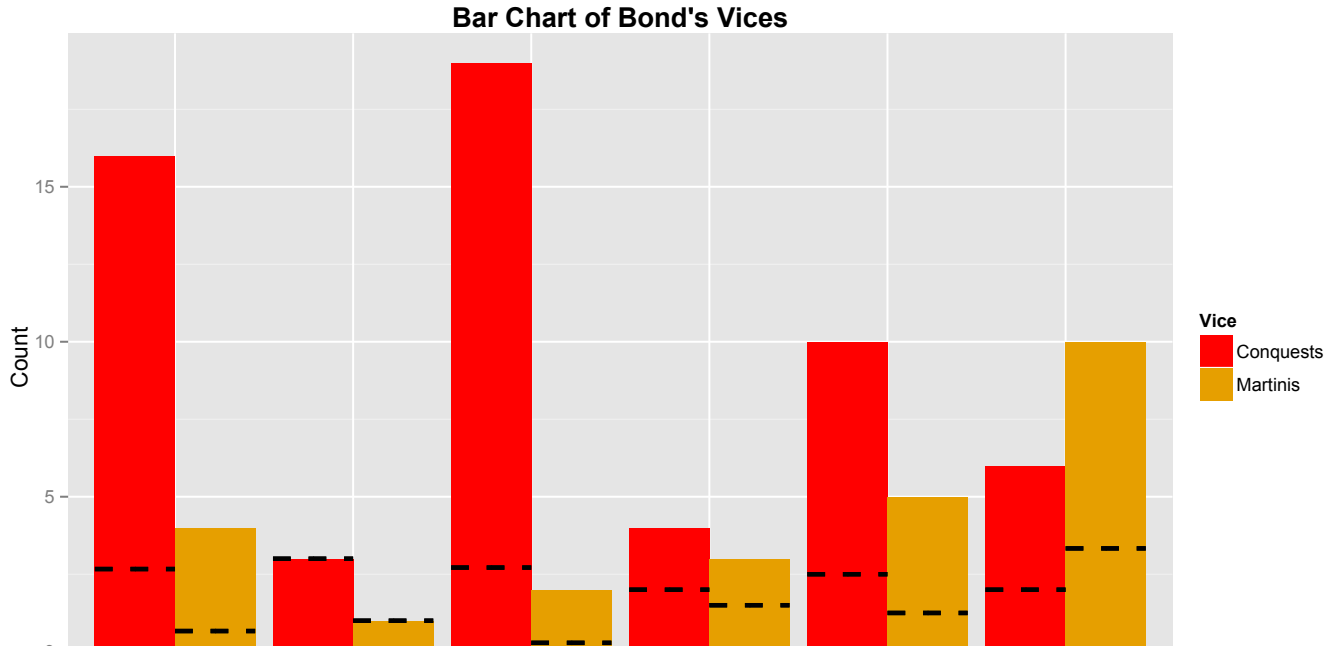


Figure 3. Bar chart showing total number of martinis and conquests by Bond actor. Dashed lines represent the mean for that actor.

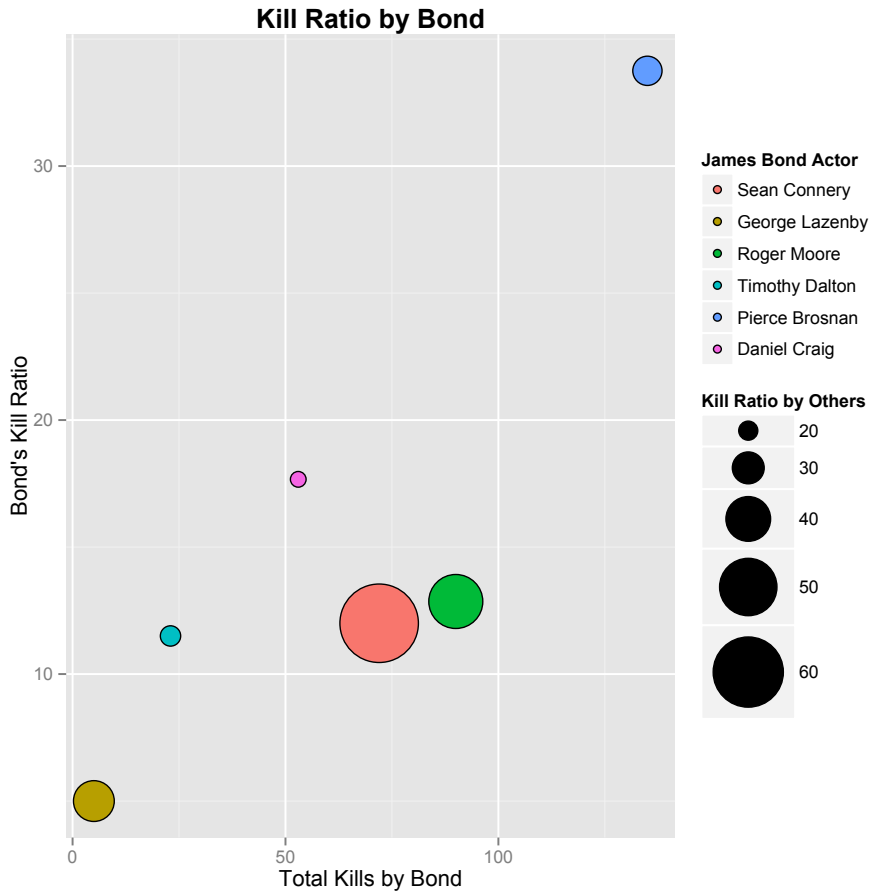


Figure 4. Bubble plot of each Bond actor's kill ratio (i.e., average kills per film) versus his total kills. The area of the bubbles is proportional to the kill ratio by others in the corresponding films.

Building a Statistical Model of the Bond Movies

The adjusted worldwide gross and average user ratings are variables that provide a good indication of each movie's success (see Figure 5). The Pearson correlation between these variables is +0.5659, which is a moderately high value and significantly different from 0 (p-value = 0.0049). Treating these variables as a bivariate response, we build a multivariate multiple linear regression model and explore the variables listed in the previous section as possible predictors in the model.

We construct the multivariate analysis of variance (MANOVA) table to assess the significance of each independent variable. We compute the following four multivariate test statistics: Wilks' lambda, Pillai's trace, the Hotelling-Lawley trace, and Roy's greatest root. Using a significance level of 0.05, we perform backwards elimination by removing the least significant variable based on the MANOVA results. Using this approach, we end up with only the Bond actor and the adjusted budget as significant predictors of the bivariate response. In fact, our decisions based on the p-values for each predictor is consistent in the four test statistics. Since our decisions do not vary across the four tests, we are confident in the results.

Our final multivariate regression model treats the adjusted worldwide gross and average rating as the response vector with Bond actor and adjusted budget as predictors. We next analyze the residual vectors for each regression to assess the normality and constant variance assumptions. The Shapiro-Wilk test for normality applied to the residuals yields p-values of

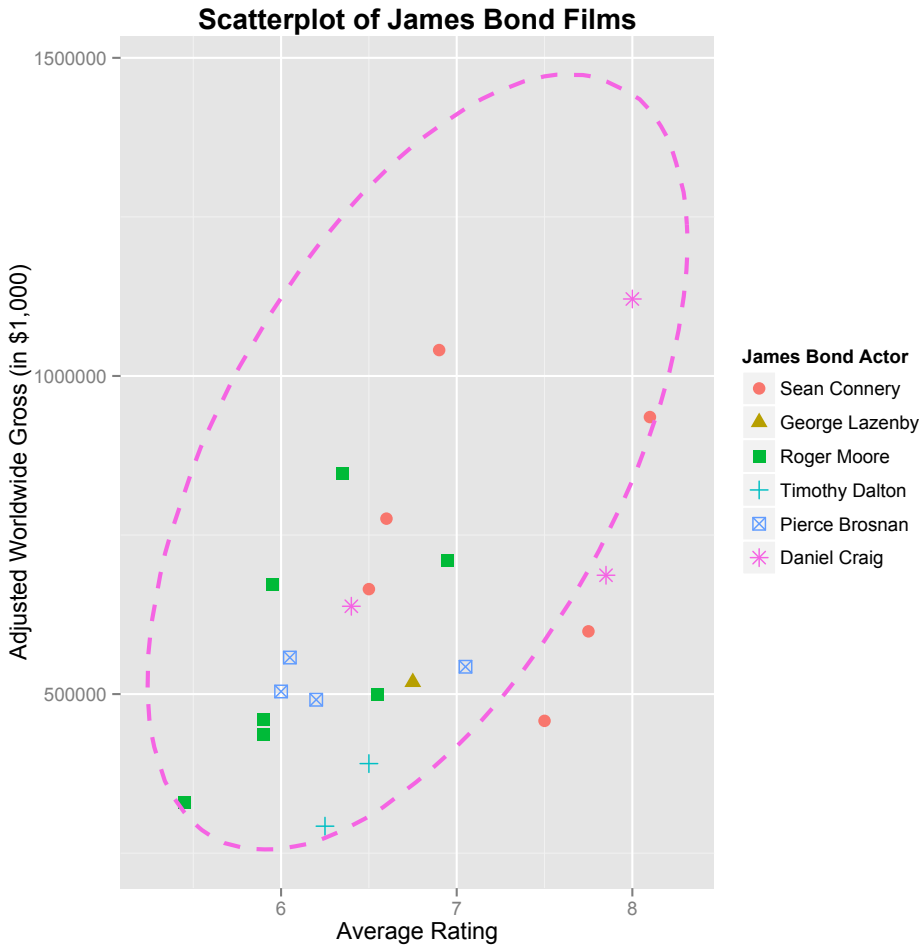


Figure 5. Scatterplot of the adjusted sales versus average rating for the James Bond films. The 95% prediction ellipse for the next James Bond film starring Daniel Craig is overlaid.

0.2780 and 0.2352 for the adjusted worldwide gross and average user relationships, respectively. The corresponding results from the Levene test for constant variance yields p-values of 0.4533 and 0.4039. Therefore, we claim that the normal regression assumptions are appropriate for this model.

We also construct a 95% prediction ellipse for the next James Bond film. It has been confirmed that Craig is signed to star in two more James Bond films, with the next film slated for release in late 2015. Using our estimated multivariate regression model and assuming that the next film will have a budget around $250 million, we obtain an estimated average rating for the film of about 6.80, while the estimated worldwide gross of the film is around $865 million. The corresponding 95% prediction ellipse is overlaid on the data plotted in Figure 5.

Ranking the Bonds

Anyone who considers himself or herself even a casual Bond fan typically opines on who they think is the best James Bond. In this section, we present a simple ranking methodology as a way to answer this question. We consider the worldwide gross, net profit of the films (i.e., the adjusted worldwide gross minus the adjusted budget), the average user ratings of the films, and the number of films featuring each actor. For each variable, we rank the James Bond actors and then calculate the means of their rankings. The

Table 2—Rankings of the James Bond Actors by Different Variables

Bond Actor	Ranking by Average Worldwide Gross	Ranking by Average Net Profit	Ranking by Average Movie Rating	Ranking by Number of Movies Made	Overall Ranking
Sean Connery	2	1	2	2	1
Daniel Craig	1	2	1	4	2
Roger Moore	3	3	6	1	3
Pierce Brosnan	4	5	5	3	4
George Lazenby	5	4	3	6	5
Timothy Dalton	6	6	4	5	6

Note: The last column is the ranking of the average of the four ranking variables.

lowest mean is given a final ranking of 1, and the highest mean is given a ranking of 6. All these rankings are reported in Table 2.

The results show that Connery comes out as the best James Bond, according to our methodology. This is not unexpected considering he defined the role and has some of the most-popular movies during his tenure as 007, like "From Russia with Love" and "Goldfinger." Craig is second, which again is not unexpected considering he is the current James Bond and his three films have been very successful. Third is Moore, who starred in the most James Bond films of any actor. While we noted the decline in the adjusted worldwide gross of his films, their success is still quite good considering they were released during years that saw competition from box office hits like "E.T." and the original "Star Wars" films.

Ranked fourth by our methodology is Brosnan, whose debut film "Goldeneye" was very well received after the franchise took a six-year hiatus. Each of his subsequent movies saw success at the box office, but they are sometimes viewed as declining in quality, as demonstrated with the decreasing ratings in Table 1. Fifth is Lazenby, who only starred in one James Bond film, "On Her Majesty's Secret Service." Fortunately, this movie is often regarded by critics and fans as being one of the best in the Bond canon as we see the character humanized by getting married, but only to have his wife killed on their wedding day. In sixth place is Timothy Dalton, who was actually asked by the original producer of the James Bond films, Cubby Broccoli, to succeed Connery back in the early 1970s. Dalton passed at the time because he felt he was "too young" for the role, only being in his late 20s at the time. While Dalton is a classically trained actor who spent time performing with the Royal Shakespeare

Company, he was unfortunate in that his films were made during a time when the Bond formula was seen as stale. In fact, he was slated to star in a third James Bond film, but the concept was shelved due to studio issues until Brosnan made his Bond debut in "Goldeneye."

Discussion

This study explored the James Bond film franchise through various data visualizations and by building a multivariate regression model that characterizes the worldwide box office receipts and average user ratings as a function of the actors and the film's budgets. From this model, we constructed a prediction region to characterize where the worldwide gross and user ratings will likely fall for the next James Bond film starring Craig. Finally, we averaged the rankings of certain variables for each James Bond actor to provide an overall ranking. Based on this ranking methodology, we determined that Connery is the "best" actor to have played 007.

While the model we presented is a way to characterize the films based on a few variables, the true success and endurance of the films is, perhaps, a topic for a more-philosophical discussion. It could be that we enjoy how the James Bond films depict travels to exotic locales and provide adventurous tales that romanticize espionage. It could be that while Bond is a flawed human being, at the core of his objective is to protect "Queen and Country" and we enjoy seeing him being the proverbial good that triumphs over evil. More personally, it could simply be that we relate to the era and style reflected in a particular Bond film that reminds us of a particular time in our own lives.

Regardless, the appeal of James Bond has spanned many generations and the films have endured through numerous global events. Even with so many uncertainties in an ever-changing world, we can rest assured that "James Bond will return…" C

Further Reading

Cork, J., and Stutz, C. 2009. *James Bond encyclopedia*. New York, USA: DK Publishing.

Desowitz, B. 2012. *James Bond unmasked*. Maryland, USA: Spies Publishing.

Dodds, K. 2003. Licensed to stereotype: Geopolitics, James Bond, and the spectre of balkanism. *Geopolitics* 8(2):125–156.

Johnson, R.A., and Wichern, D.W. 2007. *Applied multivariate statistical analysis (6th ed.)*. New Jersey, USA: Prentice Hall.

Moore, R. 2012. *Bond on Bond: Reflections on 50 years of James Bond movies*. Connecticut, USA: Lyons Press.

Neuendorf, K.A., Gore, T.D., Dalessandro, A., Janstove, P., and Snyder-Suhy, S. 2010. Shaken and stirred: A content analysis of women's portrayals in James Bond films. *Sex Roles* 62(11/12):74–77.

About the Author

Derek Young is a research mathematical statistician at the U.S. Census Bureau, lecturer in statistics at Penn State University, and Accredited Professional Statistician™. His research interests include tolerance regions, mixture models, and statistical depth functions.

How We Know that the Earth is Warming

Peter Guttorp

This article originally appeared in *CHANCE* 30.4.

There is no doubt that global temperatures are increasing, and that human greenhouse gas emissions largely are to blame, but how do we go about measuring global temperature? It is not just a matter of reading an instrument.

In Figure 1, we see a variety of curves depicting annual global mean temperature. They are not the same, although they all show a strong increase after about 1980. Different groups, using different data and different techniques, have computed the different curves. It would be hoped that the curves would all be measurements of the annual global mean temperature, but global mean temperature is not something that can be measured directly using an instrument. On the other hand, it is the quantity most commonly used to indicate global warming.

Where do the numbers come from? We will go through some issues that are associated with determining surface temperature, and illustrate some of the uses of these temperatures.

Local Daily Mean

The basic measurements that go into the calculation of global mean temperature are readings of thermometers or other instruments determining temperature. For land stations, these instruments are typically kept in some kind of box in an open, flat space covered with grass (see Figure 2). The box keeps direct sunlight from hitting the instrument but allows wind to penetrate it.

Readings are done at different schedules in different countries. The modern instruments measure continuously, but the measurements are not always recorded. In the United States, daily maximum and minimum temperature are recorded, and their average is the daily mean temperature. In Sweden, three hourly readings throughout the day are combined with the minimum and the maximum to calculate the daily mean temperature. In Iceland, linear combinations of two readings in the morning and afternoon are used.

Modern instruments can compute the daily average automatically, but to compare to historical data, a specific averaging method has to be applied.

Local Annual Mean Temperature

Once you have a daily mean temperature, it is easy to compute an annual mean temperature: Sum all the daily means and divide by

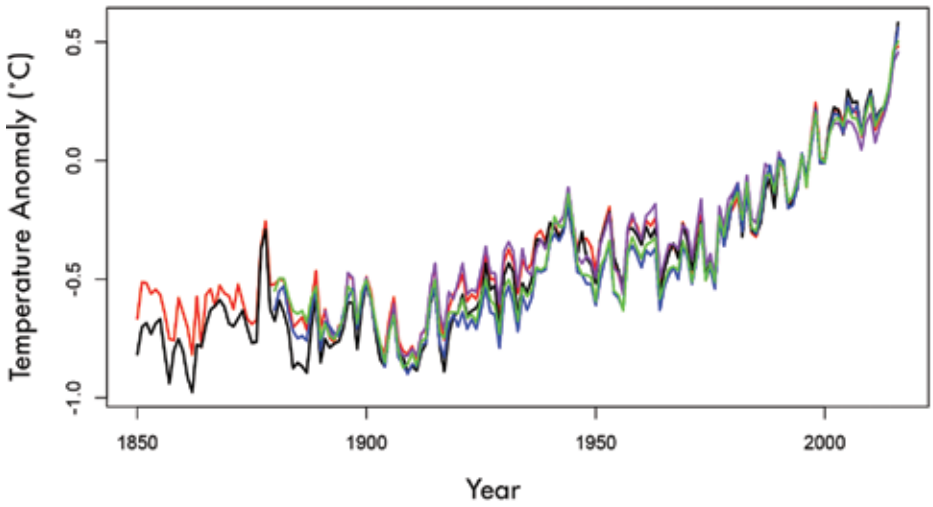


Figure 1. Five estimates of the annual global mean anomalies relative to 1981–2010: Black is from Berkeley Earth, red from the UK Met Office Hadley Center, purple from the Japanese Met Office, blue from the Goddard Institute for Space Science (GISS), and green from the National Oceanic and Atmospheric Administration (NOAA).

Figure 2. Thermometer and other instruments at Stockholm Observatory, where measurements have been made daily since 1756. The station has been moved short distances twice during this time. The box to the left is a Stephenson screen, and was used for the measurements until 2006. The pipe sticking up in the middle contains the modern measurement devise that has been used since then.
Photograph courtesy of Peter Guttorp.

the number of days in the year. What is often used instead of the annual mean is something called a mean anomaly: How much did the year deviate from the average over a reference period? This makes it easier to compare sites at different altitudes, for example. A station at a higher elevation always tends to be colder than one at a lower elevation, but anomalies allow us to see if both sites are colder than usual.

The largest collection of land station data, used in the Berkeley Earth global temperature series, has some 39,000 stations and a total of 1.6 billion temperature measurements.

Sea Surface Temperature

Since more than two-thirds of the surface of our planet is water, it is not enough to take temperature measurements on land to compute a global average. Ocean-faring ships have long kept daily logbooks, with measurements of wind, air temperature, and water temperature. The water temperature used to be taken in a bucket of seawater. Later, it would be measured at the cooling water intake for the motor. Of course, ships do not travel everywhere on the oceans and, therefore, there are fairly large areas of ocean where we have no sea surface temperature measurements from ships.

In some of these areas, there are buoys that measure the temperature. Over the last several decades, there have been satellite measurements of

The main groups estimating global mean temperature

- Hadley Center of the UK Met Office with the Climate Research Unit of the University of East Anglia, United Kingdom
- Goddard Institute for Space Science (part of NASA), USA
- National Centers for Environmental Information (part of NOAA), USA
- Japanese Met Office, Japan
- Berkeley Earth Project, USA

A simultaneous confidence band for n normally distributed estimates can be obtained by the Bonferroni inequality

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} P(E_i).$$

In fact, we want the complement: $P\left(\bigcap_{i=1}^{n} E_i^c\right) \geq 1 - \sum_{i=1}^{n} P(E_i^c)$

Let E_i be the event that the true value at time i is not covered by its (pointwise) confidence set. If we let each confidence set have level $1 - \alpha / n$, we see that the probability that all parameters (in our case, the global average temperature for each year) are covered by their respective intervals is at least $1 - n(\alpha / n) = 1 - \alpha$. The confidence band then is $t_i \pm \Phi^{-1}(1 - \alpha / n)\mathrm{se}(t_i)$ where t_i is the estimated global mean temperature for year i, $\mathrm{se}(t_i)$ is the standard error of the estimate, and Φ^{-1} is the normal quantile function (inverse of the cdf).

sea surface temperature; for over a decade, floats that measure the temperature profile of the water have been dropped all over the oceans.

The largest collection of ocean data, the ICOADS 3.0 data set, uses about 1.2 billion different records.

Combining All the Measurements

To combine the many measurements over land and oceans into an average global temperature requires estimating the temperature anomaly where there are no actual measurements, such as on a regular grid, and then averaging the estimates and measurements (if any) over the grid. Such estimation tools are derived in what is called *spatial statistics*, although atmospheric scientists have developed some methods on their own (through what they call objective analysis).

In essence, a statistician would treat the problem as one of regression, with data that are spatially dependent. The process has to take into account the fact that data are on a globe and not in the plane. The average temperature anomaly for land and ocean can be computed separately, and the global mean temperature would then be the area weighted average of the two means.

Uncertainty

There are several sources of uncertainty in the determination of global mean temperature. First of all, each measurement has error associated with it. Second, how to deal with missing areas of measurement causes uncertainty. The choice of measurement stations can also be a source of uncertainty, as can the homogenization of measurements, such as when stations are moved or measurement devices updated. There are other sources of uncertainty as well.

It is important to try to quantify the uncertainty in global mean temperature. Different groups approach this issue in different ways. Figure 3 uses the Hadley series uncertainties to compute a simultaneous Bonferroni-based 95% confidence band for global average temperature. The term simultaneous means that the confidence band covers all the true temperatures at the same time with 95% probability, as opposed to a pointwise confidence interval, which only covers the true temperature at a particular time point with 95% probability.

Ranking

In January 2017, NOAA made the claim that the global mean temperature had set a record for the third straight year. This statement is not quite accurate: For the third straight year, the *estimated* global mean temperature had set a record. In fact, four of the five series had this feature, while the Berkeley series showed 2005 as warmer than and 2010 tied with 2014. Only two of the estimates (the Hadley series

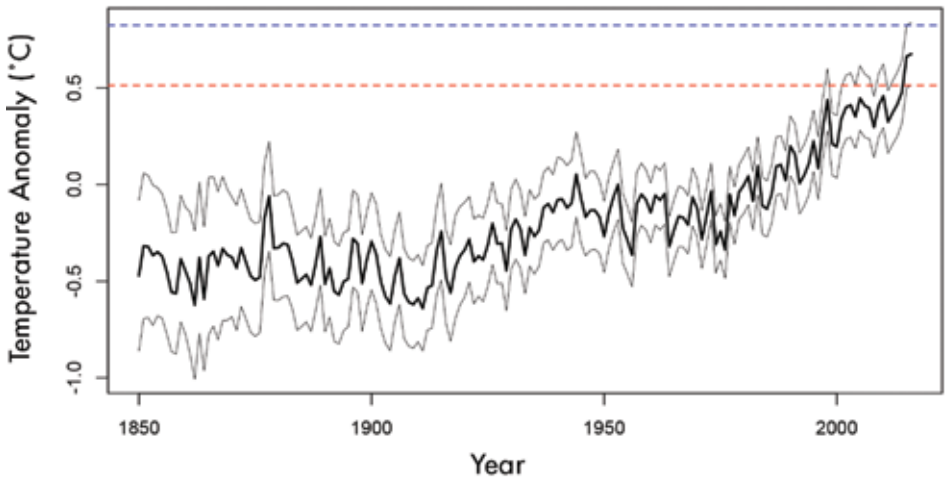


Figure 3. Hadley series with red dashed line being the lower 95% simultaneous confidence bound on the 2016 temperature and blue dashed line the upper bound on the 2015 temperature.

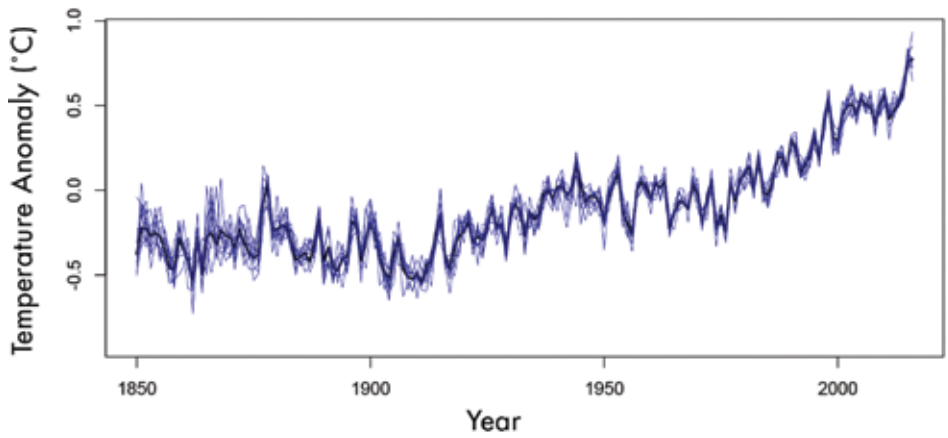


Figure 4. 10 realizations (blue) of possible Hadley temperature series and the Hadley estimate of global mean temperature (black).

and the Berkeley series) provide uncertainty estimates.

Figure 3 shows the Hadley series with associated simultaneous 95% confidence bands. If the 2015 actual temperature (which we do not know) were at the high end of its confidence band (blue dashed line), and the 2016 was at the low end of its band (red dashed line), it is quite possible that 2015 could have been substantially warmer than 2016, but that 2016 clearly was warmer than any year before 1998.

How can we say something about the uncertainty in the rankings as opposed to the estimates? One way is to simulate repeated draws from the sampling distribution of the estimates. Since we are averaging a large number of measurements, many of which are nearly uncorrelated, a central limit theorem leads us to treat the estimates as normal, with mean equal to the actual estimate and standard deviation equal to the standard error of the estimate. Figure 4 shows 10 such realizations of the Hadley temperature series.

For each of the realizations, we can calculate the rank of 2016. The distribution of that rank tells us how likely 2016 is to be the warmest year on record: It is warmest in 58% of the simulations, while 2015 is warmest in 42%. In 10,000 simulations, 2016 was as low as the eighth-warmest in one of them.

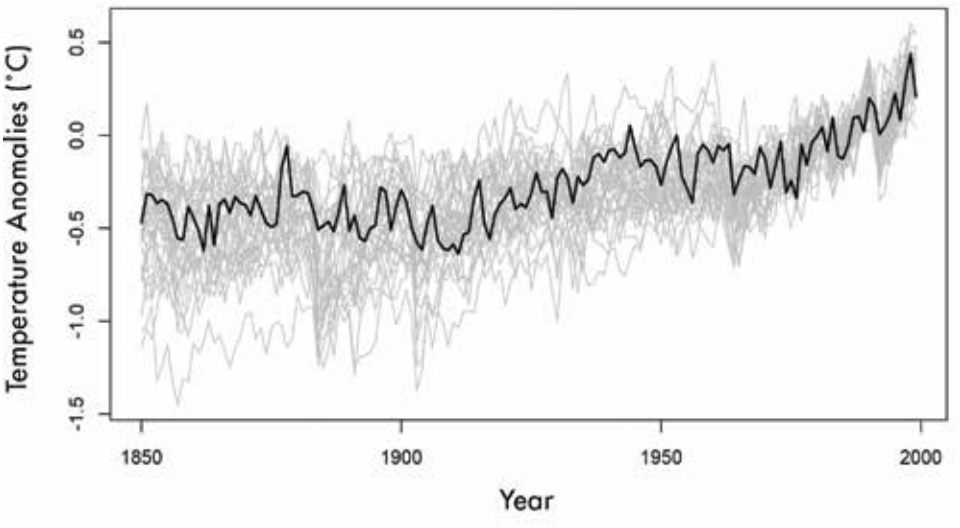


Figure 5. Global annual mean temperature anomalies from 32 CMIP5 models with historical simulations (gray), and the Hadley Center data series (black). Reference period is 1970–1999.

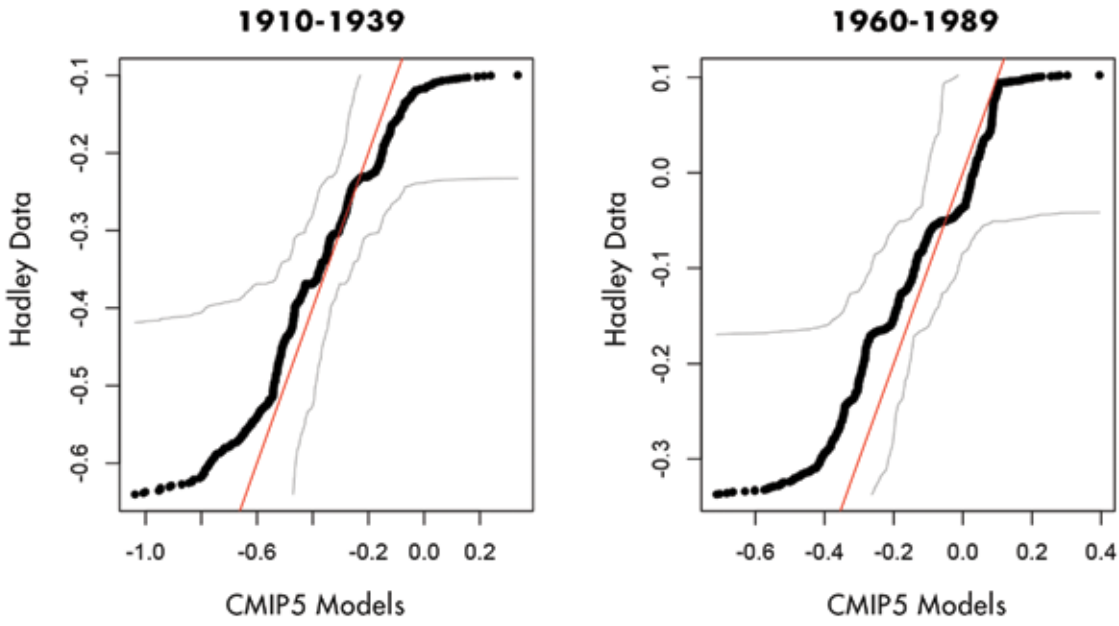


Figure 6. QQ-plots of historical climate model simulations against Hadley Center data or two 30-year periods. The gray lines are simultaneous 95% confidence bands, and the red lines indicate equal distributions.

How about all three years—2014–2016—being record-breakers? That happened in 21% of the simulations, and in the actual Hadley estimates, of course.

Models and Data

Climate, from a statistical point of view, is the distribution of weather. Climate change means that this distribution is changing over time. The World Meteorological Organization recommends using 30 years to estimate climate. This definition indicates, for example, that it does not make sense to look at shorter stretches of data to try to assess questions such as "Is global warming slowing down?"

What is a Climate Model?

A climate model is a deterministic model describing the atmosphere, sometimes the oceans, and sometimes also the biosphere. It is based on a numerical solution of coupled partial differential equations on a grid. In fact, the equations for the atmosphere are essentially the same as for weather prediction, but the latter is an initial value problem (we use today's weather to forecast tomorrow's) of a chaotic system, while the climate models has to show long-term stability. Many processes, such as hurricanes or thunderstorms, are important in transferring heat between different layers in the model, but often take place at a scale that is at most similar to a grid square, and sometimes much smaller.

Different climate models deal with this subgrid variability differently and, as a consequence, the detailed outputs are different. CMIP5 is a large collection of model runs, using the same input variables (solar radiation, volcanic eruptions, greenhouse gas concentrations, etc.). These model outputs were used for the latest IPCC report in 2013. Figure 5 shows the global mean temperature anomalies (with respect to 1970–1999) with the corresponding Hadley Center series.

Many tests have been developed to compare some aspects of distributions, such as means or medians. To compare two entire distributions, we can plot the quantiles of one against the other (called a quantile-quantile plot or QQ-plot). An advantage of this plot is that if the distributions are the same, then the plot will be a straight line. Of course, we will be estimating the quantiles from data, so there will be uncertainty. Another advantage of the QQ-plot is being able to develop simultaneous confidence bands, enabling a simple test of equal distributions: Does the line $y=x$ fit inside the confidence band?

Comparing Distributions

It is not trivial to compare climate model output to data. Remember, the climate model represents the *distribution* of the data. The observations in Figure 5 are, therefore, not directly comparable to the model runs. Instead, we need to compare the distributions of model output and data, respectively.

Figure 6 compares these distributions using QQ-plots for two 30-year stretches. In both cases, the distribution of the data fit the distribution of the ensemble of model outputs quite well, in that the red $y=x$ line falls inside the simultaneous 95% confidence bands. Since we have 32 x 30 observations of the models, and only 30 of the data, the empirical tails of the model distribution are much longer than the tails of the data, but the confidence band is quite wide in the tails, meaning that we are very uncertain there.

Thus, for these two time intervals and for the global mean temperature variable, the ensemble of CMIP5 models and the Hadley Center data seem to have the same distribution—they are describing the same climate. ◼

Further Reading

Arguez, A., Karl, T.R., Squires, M.F., and Vose, R.S. 2013. Uncertainty in annual rankings from NOAA's global temperature time series. *Geophysical Research Letters* 40:5,965–5,969.

Doksum, K. 1974. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Ann. Statist.* 2:267–277.

Guttorp, P. 2014. Statistics and Climate. *Annual Reviews of Statistics and its Applications* 1:87–101.

Katz, R.W., Craigmile, P.F., Guttorp, P., Haran, M., Sanso, B., and Stein, M.L. 2013. Uncertainty analysis in climate change assessments. *Nature Climate Change* 3:769–771.

About the Author

Peter Guttorp is a professor at the Norwegian Computing Center in Oslo and professor emeritus at the University of Washington in Seattle. He has worked on stochastic models in a variety of scientific applications, such as hydrology, climatology, and hematology. He has published six books and about 200 scientific papers.

A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks

Miguel A. Hernán, John Hsu, and Brian Healy

For much of the recent history of science, learning from data was the academic realm of statistics,[1,2] but in the early 20th century, the founders of modern statistics made a momentous decision about what could and could not be learned from data: They proclaimed that statistics could be applied to make causal inferences when using data from randomized experiments, but not when using nonexperimental (observational) data.[3,4,5] This decision classified an entire class of scientific questions in the health and social sciences as not amenable to formal quantitative inference.

Not surprisingly, many scientists ignored the statisticians' decree and continued to use observational data to study the unintended harms of medical treatments, health effects of lifestyle activities, or social impact of educational policies. Unfortunately, these scientists' causal questions often were mismatched with their statistical training. Perplexing paradoxes arose; for example, the famous "Simpson's paradox" stemmed from a failure to recognize that the choice of data analysis depends on the causal structure of the problem.[6] Mistakes occurred. For example, as a generation of medical researchers and clinicians believed that postmenopausal hormone therapy reduced the risk of heart disease because of data analyses that deviated from basic causal considerations. Even today, confusions generated by a century-old refusal to tackle causal questions explicitly are widespread in scientific research.[7]

To bridge science and data analysis, a few rogue statisticians, epidemiologists, econometricians, and computer scientists developed formal methods to quantify causal effects from observational data. Initially, each discipline emphasized different types of causal questions, developed different terminologies, and preferred different data analysis techniques. By the beginning of the 21st century, while some conceptual discrepancies remained, a unified theory of quantitative causal inference had emerged.[8,9]

We now have a historic opportunity to redefine data analysis in such a way that it naturally accommodates a science-wide framework for causal inference from observational data. A recent influx of data analysts, many not formally trained in statistical theory, bring a fresh attitude that does not a priori exclude causal questions. This new wave of data analysts refer to themselves as data scientists and to their activities as data science, a term popularized by technology companies and embraced by academic institutions.

Data science, as an umbrella term for all types of data analysis, can tear down the barriers erected by traditional statistics; put data analysis at the service of all scientific questions, including causal ones; and prevent unnecessary inferential mistakes. We may miss our chance to successfully integrate data analysis into all scientific

[1]Tukey, J.W. 1962. The future of data analysis. *Annals of Mathematical Statistics* 33:1-67.
[2]Donoho, D. 2017. 50 years of data science. *Journal of Computational and Graphical Statistics* 26(4):745–66.
[3]Pearl, J. 2009. *Causality: Models, Reasoning, and Inference* (2nd edition). New York: Cambridge University Press.
[4]Fisher, R.A. 1925. *Statistical Methods for Research Workers*, 1st ed. Edinburgh: Oliver and Boyd.
[5]Pearson, K. 1911. *The Grammar of Science*, 3rd ed. London: Adam and Charles Black.
[6]Hernán, M.A., Clayton, D., and Keiding, N. 2011. The Simpson's paradox unraveled. *International Journal of Epidemiology* 40(3):780–5.
[7]Hernán, MA. 2018. The C-word: Scientific euphemisms do not improve causal inference from observational data (with discussion). *American Journal of Public Health* 108(5): 616–9.
[8]Hernán, M.A., Robins J.M. 2018 (forthcoming). *Causal Inference*. Boca Raton: Chapman & Hall/CRC.
[9]Pearl, J. 2018. *The Book of Why*. New York: Basic Books.

questions, though, if data science ends up being defined exclusively in terms of technical[10] activities (management, processing, analysis, visualization…) without explicit consideration of the scientific tasks.

A Classification of Data Science Tasks

Data scientists often define their work as "gaining insights" or "extracting meaning" from data. These definitions are too vague to characterize the scientific uses of data science. Only by precisely classifying the "insights" and "meaning" that data can provide will we be able to think systematically about the types of data, assumptions, and analytics that are needed. The scientific contributions of data science can be organized into three classes of tasks: description, prediction, and counterfactual prediction (see table for examples of research questions for each of these tasks).

Description is using data to provide a quantitative summary of certain features of the world. Descriptive tasks include, for example, computing the proportion of individuals with diabetes in a large healthcare database and representing social networks in a community. The analytics employed for description range from elementary calculations (a mean or a proportion) to sophisticated techniques such as unsupervised learning algorithms (cluster analysis) and clever data visualizations.

Prediction is using data to map some features of the world (the inputs) to other features of the world (the outputs). Prediction often starts with simple tasks (quantifying the association between albumin levels at admission and death within one week among patients in the intensive care unit) and then progresses to more-complex ones (using hundreds of variables measured at admission to predict which patients are more likely to die within one week). The analytics employed for prediction range from elementary calculations (a correlation coefficient or a risk difference) to sophisticated pattern recognition methods and supervised learning algorithms that can be used as classifiers (random forests, neural networks) or predict the joint distribution of multiple variables.

Counterfactual prediction is using data to predict certain features of the world as if the world had been different, which is required in *causal inference* applications. An example of causal inference is the estimation of the mortality rate that would have been observed if all individuals in a study population had received screening for colorectal cancer vs. if they had not received screening.

The analytics employed for causal inference range from elementary calculations in randomized experiments with no loss to follow-up and perfect adherence (the difference in mortality rates between the screened and the unscreened) to complex implementations of g-methods in observational studies with treatment-confounder feedback (the plug-in g-formula).[11]

Note that, contrary to some computer scientists' belief, "causal inference" and "reinforcement learning" are not synonyms. Reinforcement learning is a technique that, in some simple settings, leads to sound causal inference. However, reinforcement learning is insufficient for causal inference in complex settings (discussed below).

Statistical inference is often required for all three tasks. For example, one might want to add 95% confidence intervals for descriptive, predictive, or causal estimates involving samples of target populations.

As in most attempts at classification, the boundaries between the above categories are not always sharp. However, this trichotomy provides a useful starting point to discuss the data requirements, assumptions, and analytics necessary to successfully perform each task of data science. A similar taxonomy has traditionally been taught by data scientists from many disciplines, including epidemiology, biostatistics,[12] economics,[13] and political science.[14] Some methodologists have referred to the causal inference task as "explanation,"[15] but this is a somewhat-misleading term because causal effects may be quantified while remaining unexplained (randomized trials identify causal effects even if the causal mechanisms that explain them are unknown).

Sciences are primarily defined by their questions rather than by

[10]Cleveland, W. 2001. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *Intenational Statistical Review* 69(1):21-6.
[11]Robins, J.M. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—Application to the healthy worker survivor effect. *Mathematical Modelling* 7:1,393–512 (1987. errata, Mathematicall Modelling 14:917–21).
[12]Vittinghoff, E., Glidden, D.V., Shiboski, S.C., and McCulloch, C.E. 2012. *Regression Methods in Biostatistics*. New York: Springer.
[13]Mullainathan, S., and Spiess, J. 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives* 31(2):87–106.
[14]Toshkov, D. 2016. *Research Design in Political Science*. London: Palgrave McMillan.
[15]Schmueli, G. 2010. To explain or to predict? *Statistical Science* 25(3):289–310.

their tools: We define astrophysics as the discipline that learns the composition of the stars, not as the discipline that uses the spectroscope. Similarly, data science is the discipline that describes, predicts, and makes causal inferences (or, more generally, counterfactual predictions), not the discipline that uses machine learning algorithms or other technical tools. Of course data science certainly benefits from the development of tools for the acquisition, storage, integration, access, and processing of data, as well as from the development of scalable and parallelizable analytics. This data engineering powers the scientific tasks of data science.

Prediction vs. Causal Inference

Data science has excelled at commercial applications, such as shopping and movie recommendations, credit rating, stock trading algorithms, and advertisement placement. Some data scientists have transferred their skills to scientific research with biomedical applications such as Google's algorithm to diagnose diabetic retinopathy[16] (after 54 ophthalmologists classified more than 120,000 images), Microsoft's algorithm to predict pancreatic cancer months before its usual diagnosis[17] (using the online search histories of 3,000 users who were later diagnosed with cancer), and Facebook's algorithm to detect users who may be suicidal[18] (based on posts and live videos).

All these applications of data science have one thing in common: They are predictive, not causal. They map inputs (an image of a human retina) to outputs (a diagnosis of retinopathy), but they do not consider how the world would look like under different courses of action (whether the diagnosis would change if we operated on the retina).

Mapping observed inputs to observed outputs is a natural candidate for automated data analysis because this task only requires: 1) a large data set with inputs and outputs, 2) an algorithm that establishes a mapping between inputs and outputs, and 3) a metric to assess the performance of the mapping, often based on a gold standard.[19] Once these three elements are in place, as in the retinopathy example, predictive tasks can be automated via data-driven analytics that evaluate and iteratively improve the mapping between inputs and outputs without human intervention.

More precisely, the component of prediction tasks that can be automated easily is the one that does not involve any expert knowledge. Prediction tasks require expert knowledge to specify the scientific question—what to input and what outputs—and to identify/generate relevant data sources.[20] (The extent of expert knowledge varies with different prediction tasks.[21]) However, no expert knowledge is required for prediction after candidate inputs and the outputs are specified and measured in the population of interest. At this point, a machine learning algorithm can take over the data analysis to deliver a mapping and quantify its performance. The resulting mapping may be opaque, as in many deep learning applications, but its ability to map the inputs to the outputs with a known accuracy in the studied population is not in question.

The role of expert knowledge is the key difference between prediction and causal inference tasks. Causal inference tasks require expert knowledge not only to specify the question (the causal effect of what treatment on what outcome) and identify/generate relevant data sources, but also to describe the causal structure of the system under study. Causal knowledge, usually in the form of unverifiable assumptions,[22,23] is necessary to guide the data analysis and to provide a justification for endowing the resulting numerical estimates with a causal interpretation. In other words, the validity of causal inferences depends on structural knowledge, which is usually incomplete, to supplement the information in the data. As a consequence, no algorithm

[16]Gulshan, V., Peng, L., Coram, M., et al. 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 316(22):2,402–10.

[17]Paparrizos, J., White, R.W., and Horvitz, E. 2016. Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results. *Journal of Oncological Practice* 12(8):737–44.

[18]Rosen, G. 2017. Getting Our Community Help in Real Time. *https://newsroom.fb.com/news/2017/11/getting-our-community-help-in-real-time/* (accessed April 26, 2018).

[19]Brynjolfsson, E., and Mitchell, T. 2017. What can machine learning do? Workforce implications. *Science* 358(6370):1,530–4.

[20]Conway, D. 2010. The Data Science Venn Diagram. Accessed October 9, 2018. *http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.*

[21]Beam, A.L., and Kohane I.S. 2018. Big Data and Machine Learning in Health Care. *JAMA* 319(13):1,317–8.

[22]Robins, J.M. 2001, Data, design, and background knowledge in etiologic inference. *Epidemiology* 11:313–20.

[23]Robins, J.M., and Greenland, S. 1986. The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology* 123(3):392–402.

can quantify the accuracy of causal inferences from observational data. The following simplified example helps fix ideas about the different role of expert knowledge for prediction versus causal inference.

Example

Suppose we want to use a large health records database to predict infant mortality (the output) using clinical and lifestyle factors collected during pregnancy (the inputs). We have just applied our expert knowledge to decide what the output and candidate inputs are, and to select a particular database in the population of interest. The only requirement is that the potential inputs must precede the outputs temporally, regardless of the causal structure linking them. At this point of the process, our expert knowledge will not be needed any more: An algorithm can provide a mapping between inputs and outputs at least as good as any mapping we could propose and, in many cases, astoundingly better.

Now suppose we want to use the same health records database to determine the causal effect of maternal smoking during pregnancy on the risk of infant mortality. A key problem is confounding: Pregnant women who do and do not smoke differ in many characteristics (including alcohol consumption, diet, access to adequate prenatal care) that affect the risk of infant mortality. Therefore, a causal analysis must identify and adjust for those confounding factors which, by definition, are associated with both maternal smoking and infant mortality.

However, not all factors associated with maternal smoking and infant mortality are confounders that should be adjusted for. For example, birthweight is strongly associated with both maternal smoking and infant mortality, but adjustment for birthweight induces bias because birthweight is a risk factor that is itself causally affected by maternal smoking. In fact, adjustment for birthweight results in a bias often referred to as the "birthweight paradox": Low birthweight babies from mothers who smoked during pregnancy have a lower mortality than those from mothers who did not smoke during pregnancy.[24]

An algorithm devoid of causal expert knowledge will rely exclusively on the associations found in the data and is therefore at risk of selecting features, like birthweight, that increase bias. The "birthweight paradox" is indeed an example of how the use of automatic adjustment procedures may lead to an incorrect causal conclusion. In contrast, a human expert can readily identify many variables that, like birthweight, should not be adjusted for because of their position in the causal structure.

A human expert also may identify features that should be adjusted for, even if they are not available in the data, and propose sensitivity analyses[25] to assess the reliability of causal inferences in the absence of those features. In contrast, an algorithm that ignores the causal structure will not issue an alert about the need to adjust for features that are not in the data.

Given the central role of (potentially fallible) expert causal knowledge in causal inference, it is not surprising that researchers look for procedures to alleviate the reliance of causal inferences on causal knowledge. Randomization is the best such procedure.

When a treatment is randomly assigned, we can unbiasedly estimate the average causal effect of treatment assignment *in the absence of detailed causal knowledge about the system under study*. Randomized experiments are central in many areas of science where relatively simple causal questions are asked.[26] Randomized experiments are also commonly used, often under the name A/B testing, to answer simple causal questions in commercial web applications. However, randomized designs are often infeasible, untimely, or unethical in the extremely complex systems studied by health and social scientists.[26]

A failure to grasp the different role of expert knowledge in prediction and causal inference is a common source of confusion in data science (the confusion is compounded by the fact that predictive analytic techniques, such as regression, can also be used for causal inference when combined with causal knowledge).

Both prediction and causal inference require expert knowledge to formulate the scientific question i, but only causal inference requires causal expert knowledge to answer the question. As a result,

[24]Hernández-Díaz, S., Schisterman, E.F., and Hernán, M.A. 2006. The birth weight "paradox" uncovered? *American Journal of Epidemiology* 164(11):1,115–20.

[25]Robins, J.M., Rotnitzky, A., and Scharfstein, D.O. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In Halloran E., and Berry D., eds. *Statistical Methods in Epidemiology: The Environment and Clinical Trials*. New York: Springer Verlag; 1999:1–92.

[26]Hernán, M.A. 2015. Invited commentary: Agent-based models for causal inference-reweighting data and theory in epidemiology. *American Journal of Epidemiology* 2181(2):103–5.

the accuracy of causal estimates cannot be assessed by using metrics computed from the data, even if the data were perfectly measured in the population of interest.

Implications for Decision-making

A goal of data science is to help people make better decisions. For example, in health settings, the goal is to help decision-makers—patients, clinicians, policy-makers, public health officers, regulators—decide among several possible strategies. Frequently, the ability of data science to improve decision-making is predicated on the basis of its success at prediction.

However, the premise that predictive algorithms will lead to better decisions is questionable. An algorithm that excels at using data about patients with heart failure to predict who will die within the next five years is agnostic about how to reduce mortality. For example, a prior hospitalization may be identified as a useful predictor of mortality, but nobody would suggest that we stop hospitalizing people to reduce mortality. Identifying patients with bad prognoses is very different from identifying the best course of action for preventing or treating a disease. Worse, predictive algorithms, when incorrectly used for causal inference, may lead to incorrect confounder adjustment and therefore conclude, for example, that maternal smoking appears to be beneficial for low birthweight babies.

Predictive algorithms inform us that decisions have to be made, but they cannot help us make the decisions. For example, a predictive algorithm that identifies patients with severe heart failure does not provide information about whether heart transplant is the best treatment option. In contrast, causal analyses are designed to help us make decisions because they tackle "what if" questions. A causal analysis will, for instance, compare the benefit-risk profile of heart transplant versus medical treatment in patients with certain severity of heart failure.

Interestingly, the distinction between prediction and causal inference (counterfactual prediction) becomes unnecessary for decision-making when the relevant expert knowledge can readily be encoded and incorporated into the algorithms. A purely predictive algorithm that learns to play Go can perfectly predict the counterfactual state of the game under different moves, and a predictive algorithm that learns to drive a car can accurately predict the counterfactual state of the car if, say, the brakes are not operated.

Because these systems are governed by a set of known game rules (in the case of games like Go) or physical laws with some stochastic components (in the case of engineering applications like self-driving cars), an algorithm can eventually predict the behavior of the entire system under a hypothetical intervention.

Take the game of Go, which has been mastered by an algorithm "without human knowledge."[27] When making a move, the algorithm has access to all information that matters: game rules, current board position, and future outcomes fully determined by the sequence of moves. Further, a reinforcement learning algorithm can collect an arbitrary amount of data by playing more games (conducting numerous experiments), which allows it to learn by trial and error. In this setting, a cleverly designed algorithm running on a powerful computer can spectacularly outperform humans—but this form of causal inference has, at this time in history, a restricted domain of applicability.

Many scientists work on complex systems with partly known and nondeterministic governing laws (the "rules of the game"), with uncertainty about whether all necessary data are available, and for which learning by trial and error—or even conducting a single experiment—is impossible. Even when the laws are known and the data available, the system may still be too chaotic for exact long-term prediction. For example, it was impossible to predict when and where the Chinese space station,[28] while in orbit at an altitude of about 250 km, would fall to Earth.

Consider a causal question about the effect of different epoetin strategies on the mortality of patients with renal disease. We do not understand the causal structure by which molecular, cellular, individual, social, and environmental factors regulate the effect of epoetin dose on mortality risk. As a result, it is currently impossible to construct a predictive model based on electronic health records to reproduce the behavior of the system under a hypothetical intervention on an individual. Some widely publicized disappointments in causal applications of data science, like "Watson for Oncology,"

[27]Silver, D., Schrittwieser, J., and Simonyan, K., et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550(7676):354–9.
[28]The Data Team. 2018. An out-of-control Chinese space station will soon fall to Earth. *The Economist* March 19, 2018.

Table 1—Examples of Tasks Conducted by Data Scientists Working with Electronic Health Records

	Data Science Task		
	Description	**Prediction**	**Causal inference**
Example of scientific question	How can women aged 60–80 years with stroke history be partitioned in classes defined by their characteristics?	What is the probability of having a stroke next year for women with certain characteristics?	Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics?
Data	• Eligibility criteria • Features (symptoms, clinical parameters …)	• Eligibility criteria • Output (diagnosis of stroke over the next year) • Inputs (age, blood pressure, history of stroke, diabetes at baseline)	• Eligibility criteria • Outcome (diagnosis of stroke over the next year) • Treatment (initiation of statins at baseline) • Confounders • Effect modifiers (optional)
Examples of analytics	Cluster analysis …	Regression Decision trees Random forests Support vector machines Neural networks …	Regression Matching Inverse probability weighting G-formula G-estimation Instrumental variable estimation …

have arguably resulted from trying to predict a complex system that is still poorly understood and for which a sound model to combine expert causal knowledge with the available data is lacking.[29]

The striking contrast between the cautious attitude of most traditional data scientists (statisticians, epidemiologists, economists, political scientists…) and the "can do" attitude of many computer scientists, informaticians, and others seems to be, to a large extent, the consequence of the different complexity of the causal questions historically tackled by each of these groups. Epidemiologists and other data scientists working with extremely complex systems tend to focus on the relatively modest goal of designing observational analyses to answer narrow causal questions about the average causal effect of a variable (such as epoetin treatment), rather than try to explain the causal structure of the entire system or identify globally optimal decision-making strategies.

On the other hand, newcomers to data science have often focused on systems governed by known laws (like board games or self-driving cars), so it is not surprising that they have deemphasized the distinction between prediction and causal inference. Bringing this distinction to the forefront is, however, urgent as an increasing number of data scientists address the causal questions traditionally asked by health and social scientists. Sophisticated prediction algorithms may suffice to develop unbeatable Go software and, eventually, safe self-driving vehicles, but causal inferences in complex systems (say, the effects of clinical strategies to

[29]Ross, C., and Swetlitz, I. 2017. IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. *STAT*. *https://www.statnews.com/2017/09/05/watson-ibm-cancer/*.

[30]Pearl, J. 2018. Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. *Technical Report* R-475 (*http://ftp.cs.ucla.edu/pub/stat_ser/r475.pdf*. Accessed April 26, 2018.

treat a chronic disease) need to rely on data analysis methods equipped with causal knowledge.[30]

Processes and Implications for Teaching

The training of data scientists tends to emphasize mastering tools for data management and data analysis. While learning to use these tools will continue to play a central role, it is important that the technical training of data scientists makes it clear that the tools are at the service of distinct scientific tasks—description, prediction, and causal inference.

A training program in data science can, therefore, be organized explicitly in three components, each devoted to one of the three tasks of data science. Each component would describe how to articulate scientific questions, data requirements, threats to validity, data analysis techniques, and the role of expert knowledge (separately for description, prediction, and causal nference). This is the approach that we adopted to develop the curriculum of the Clinical Data Science core at the Harvard Medical School, which three cohorts of clinical investigators have now learned.

Our students first learn to differentiate between the three tasks of data science, then how to generate and analyze data for each task, as well as the differences between tasks. They learn that description and prediction may be affected by selection and measurement biases, but that only causal inference is affected by confounding. After learning predictive algorithms, teams of students compete against each other in a machine learning competition to develop the best predictive model (in an application of the Common Task Framework[2]).

By contrast, after learning causal inference techniques, students understand that a similar competition is not possible because their causal estimates cannot be ranked automatically. Teams with different subject-matter knowledge may produce different causal estimates, and there often is no objective way to determine which one is closest to the truth using the existing data.[31]

Then students learn to ask causal questions in terms of a contrast of interventions conducted over a fixed time period as would be specified in the protocol of a (possibly hypothetical) experiment, which is the target of inference.

For example, to compare the mortality under various epoetin dosing strategies in patients with renal failure, students use subject-matter knowledge to 1) outline the design of the hypothetical randomized experiment that would estimate the causal effect of interest—the target trial, 2) identify an observational database with sufficient information to approximately emulate the target trial, and 3) emulate the target trial and therefore estimate the causal effect of interest using the observational database. We discuss why causal questions that cannot be translated into target experiments are not sufficiently well-defined,[31] and why the accuracy of causal answers cannot be quantified using observational data. In parallel, the students also learn computer coding and the basics of statistical inference to deal with the uncertainty inherent to any data analyses involving description, prediction, or causal inference.

A data science curriculum along the three dimensions of description, prediction, and causal inference facilitates interdisciplinary integration. Learning from data requires paying attention to the different emphases, questions, and analytic methods developed over several decades in statistics, epidemiology, econometrics, computer science, and others. Data scientists without subject-matter knowledge cannot conduct causal analyses in isolation: They don't know how to articulate the questions (what the target experiment is) and they don't know how to answer them (how to emulate the target experiment).

Conclusion

Data science is a component of many sciences, including the health and social ones. Therefore, the tasks of data science are the tasks of those sciences—description, prediction, causal inference. A sometimes-overlooked point is that a successful data science requires not only good data and algorithms, but also domain knowledge (including causal knowledge) from its parent sciences.

The current rebirth of data science is an opportunity to rethink data analysis free of the historical constraints imposed by traditional statistics, which have left scientists ill-equipped to handle causal questions. While the clout of statistics in scientific training and publishing impeded the introduction of a unified formal framework for causal inference in data

[31]Hernán, MA. 2019 (in press). Spherical cows in a vacuum: Data analysis competitions for causal inference. *Statistical Science*.

analysis, the coining of the term "data science" and the recent influx of "data scientists" interested in causal analyses provides a once-in-a-generation chance of integrating all scientific questions, including causal ones, in a principled data analysis framework. An integrated data science curriculum can present a coherent conceptual framework that fosters understanding and collaboration between data analysts and domain experts.

On the other hand, if the definitions of data science currently discussed in mainstream statistics take hold, causal inference from observational data will be once more marginalized, leaving health and social scientists on their own. The American Statistical Association statement on "The Role of Statistics in Data Science" (August 8, 2015) makes no reference to causal inference. A recent assessment of data science and statistics[2] did not include the word "causal" (except when mentioning the title of the course "Experiments and Causal Inference"). Heavily influenced by statisticians, many medical editors actively suppress the term "causal" from their publications.[33]

A data science that embraces causal inference must (1) develop methods for the integration of sophisticated analytics with expert causal expertise, and (2) acknowledge that, unlike for prediction, the assessment of the validity of causal inferences cannot be exclusively data-driven because the validity of causal inferences also depends on the adequacy of expert causal knowledge. Causal directed acyclic graphs[34,35] may play an important role in the development of analytic methods that integrate learning algorithms and subject-matter knowledge. These graphs can be used to represent different sets of causal structures that are compatible with existing causal knowledge and thus to explore the impact of causal uncertainty on the effect estimates.

Large amounts of data could make expert knowledge irrelevant for prediction and for relatively simple causal inferences involving games and some engineering applications, but expert causal knowledge is necessary to formulate and answer causal questions in more-complex systems. Affirming causal inference as a legitimate scientific pursuit is the first step in transforming data science into a reliable tool to guide decision-making.

Finally, the distinction between prediction and causal inference is also crucial to defining artificial intelligence (AI). Some data scientists argue that "the essence of intelligence is the ability to predict," and therefore that good predictive algorithms are a form of AI. From this point of view, large chunks of data science can be rebranded as AI (and that is exactly what the tech industry is doing). However, mapping observed inputs to observed outputs barely qualifies as intelligence. Rather, a hallmark of intelligence is the ability to predict *counterfactually* how the world would change under different actions by integrating expert knowledge and mapping algorithms. No AI will be worthy of the name without causal inference. C

About the Authors

Miguel Hernán conducts research to learn what works for the treatment and prevention of cancer, cardiovascular disease, and HIV infection. With his collaborators, he designs analyses of healthcare databases, epidemiologic studies, and randomized trials. He teaches clinical data science at the Harvard Medical School, clinical epidemiology at the Harvard-MIT Division of Health Sciences and Technology, and causal inference methodology at the Harvard T.H. Chan School of Public Health, where he is the Kolokotrones Professor of Biostatistics and Epidemiology.

John Hsu is director of the Program for Clinical Economics and Policy Analysis in the Mongan Institute, Massachusetts General Hospital, and Harvard Medical School. He studies innovations in healthcare financing and delivery, and their effects on medical quality and efficiency. He primarily uses large automated and electronic health record data sets, often exploiting natural experiments from both clinical and behavioral economics perspectives.

Brian Healy is an assistant professor of neurology at the Harvard Medical School and an assistant professor in the Department of Biostatistics at the Harvard T.H. Chan School of Public Health. He is the primary biostatistician for the Partners MS Center at Brigham and Women's Hospital and a member of the Massachusetts General Hospital (MGH) Biostatistics Center. He teaches introductory statistics in several programs and co-directs the clinical data science sequence in the master of medical science and clinical investigation with Miguel Hernán.

[32]Ruich, P. 2017. The Use of Cause-and-Effect Language in the JAMA Network Journals. *AMA Style Insider. http://amastyleinsider.com/2017/09/19/use-cause-effect-language-jama-network-journals/*. Accessed May 25, 2018.

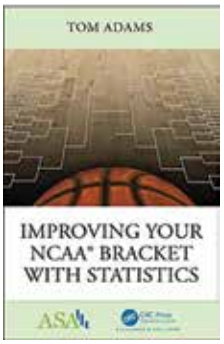
[33]Hernán, M.A, Hernández-Díaz, S., Werler, M.M., and Mitchell, A.A. 2002. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American Journal of Epidemiology* 155:176–84.

[34]Greenland, S., Pearl, J, and Robins, J.M. Causal diagrams for epidemiologic research. *Epidemiology* 1999; 10(1):37–48.

Book Excerpt: *Improving Your NCAA® Bracket with Statistics*

Tom Adams

2019. CRC Press LLC/ASA.
Reprinted with permission. Edited for house style.



Statistical Hypothesis Testing

Let's say you have been entering one bracket in a winner-take-all pool with 30 entries and a $10 entry fee for the last 10 years. You have won in one of those years. That puts you well into the black, since you have won $300 and all your fees add up to only $100. But does your success actually constitute convincing evidence that you have a superior bracket strategy?

If you were an average player, you would have a 1/30 chance of winning. (We are assuming that the pool has a rule for breaking ties, so we can ignore those.) Your chance of losing is 29/30. The outcomes from year to year are presumably independent events, so you can multiply their probabilities together. Your chance of losing in all 10 years is $(29/30)^{10}$, which comes out to about 0.71 or 71%. Your chance of winning one or more pools is about 29%, a little less than one chance in three, even if you have no superior skill. Winning only once is not that impressive. This or better would happen by chance about one-third of the time even if you are just an average player.

This reasoning constitutes a *statistical hypothesis test*. In the terminology of statistical hypothesis testing, we have failed to reject the hypothesis that you are an average player in this pool. This hypothesis (the hypothesis that you would like to reject) is called the *null hypothesis*.

You might want to claim that this test proves nothing—that you are indeed a superior player. This is called the *alternative hypothesis*. If you are right, then we failed to reject a false null hypothesis. This is called a *type II error* or a false-negative. (The *type I error* or falsepositive is when we reject a true null hypothesis.)

Statistical Power

Perhaps you could silence the scoffers via another hypothesis test. You could do another 10-year test. If you played one entry for the next 10 years, maybe you would win more than once. If you won more than once in the next 10 years, would that allow us to reject the null hypothesis?

The math is a bit more complicated for this case, but the calculations can be performed using the binomial test. The *binomial test* is appropriate because the average player's wins can be modeled as a binomial probability distribution. The binomial distribution represents the probability of any number of successes for a fixed number of trials and a fixed probability of success. In our case, we have 10 trials (pools) that result in either a win or a loss, and the win probability of the average player is 1/30.

The binomial test shows that an average player in a pool with 30 entries would have a 0.042 probability of winning two or more times in 10 years. That would be pretty convincing evidence that you are a superior pool player by the usual standards of hypothesis testing. Anything below a *significance level* of 5% is typically considered to be a good basis for rejecting the null hypothesis, but it also depends on how credible the alternative hypothesis is in the first place.

If you are picking your teams based on which mascot is the strongest, or if you claim to have extra sensory perception, then people might still think you were just lucky. If you were using bracket improvement strategies based on the peer-reviewed research, then the notion that you are a superior pool player would already be plausible.

What are the chances that you will win twice in the next 10 years? We don't have enough information

to calculate that. How much better are you than the average player? We have to know the answer to that question, or at least have an estimate. If you are using the strategies from the peer-reviewed literature, then it's reasonable to estimate that your win probability is around three times better than that of the average player. This quantity (three times better or 300%) is a measure of the *effect size* of a strategy. The inverse (1/3) is the *relative risk* of losing and is also called the *risk ratio*.

A good strategy protects against the risk of losing. Relative risk is one of many different metrics that may be used to quantify effect sizes. A particular effect-size metric is chosen based on its usefulness for communicating the magnitude of the effect or improving the experimental design. In our example, relative risk is useful for both purposes.

The average player wins 1 in 30 pools on average. If you are three times better, then you will win 1 in 10 pools or 10% of the time on average. Using the binomial test calculation, your chances of winning two or more pools over 10 years is about 0.26 or 26% of the time. The 0.26 value is called the *power* of the statistical hypothesis test. The power of a statistical test is the probability that the test will result in the rejection of the null hypothesis.

This all means that you will wait 10 years for a new hypothesis test to complete, and there is only a 26% chance that it will reject the notion that you are just average. And that is even assuming that you are three times better than average!

The problem is that the proposed statistical test is weak. How can we make it stronger? One thing to try is to enter more brackets. Let's see what would happen if you entered 15 brackets in a pool with 30 opposing brackets. Of course, your pool might have a rule that limits the number of brackets that you can enter, or the pool czar might balk when you try to do this even if there is no explicit rule against it, but you could always do a "what if?" experiment where you prepare 15 brackets before betting closes and see how they would have done if you had been able to submit them.

The null hypothesis will now be that your 15 brackets are average overall. There will be 30 opposing entries and 15 of your entries, for a total of 45 entries. You will be betting 1/3 of the entries. If the null hypothesis is true, then you have a 1/3 chance of winning the pool. According to the binomial test calculation, the chance that you will win in six or more of the 10 years is 7.6%, so that would not be good enough to reject the null hypothesis at the 5% significance level. You'd have to win seven or more times in 10 years. The probability of that happening under the null hypothesis is less than 2%.

What is the power of this test? If you win three times better than average, then you will win in all 10 years, but that is not a realistic assumption. You have been submitting the best single bracket that you could estimate, so those other 14 brackets have to be less likely to win. We need a new estimate for the effect size of betting 15 brackets instead of just one bracket.

Let's assume that 200% or two times better than average is a reasonable estimate of the effect size. In that scenario, 15 average brackets have a 1/3 chance of winning, so your superior set of 15 brackets will have a 2/3 chance of winning. If you have a 2/3 chance of winning, then your chance of winning the required seven or more pools is about 56%, according to the binomial test calculation. The power of the test is 0.56.

A statistical hypothesis test where you submit one bracket has a power of 26%—but if you submit 15 brackets, the power goes up to 56%. The 56% is not a slam dunk, but it's lot better than 26%. Using some basic statistical tools and a few assumptions, we have come up with a better design for a statistical hypothesis test.

The rest of this chapter describes an actual statistical hypothesis test using 10 years of pool data from a standard scoring pool. This pool has four payouts. The binomial test is not applicable to this data because the binomial test is only applicable to binary outcomes.

The outcome of a pool with four payouts is not just a win-lose proposition. The returns for someone playing 1/3 of the total entries for 10 years in a pool with four payouts are roughly normally distributed, so we can invoke the central limit theorem and use a statistical test that assumes normal distributions, but first we need to come up with an effective multiple-entry strategy.

Back-testing a Multiple-entry Strategy

Strategies for submitting multiple entries had been little discussed in the peer-reviewed literature. Breiter and Carlin referred to the "time-honored method of multiple pool entries." They suggested the heuristic of using multiple entries from different strategies for optimizing bracket for submission to pools with upset incentives. Clair and Letscher pointed out that submitting multiple entries could reduce dependence on a specific tournament outcome model, and Breiter and Carlin said that the variability in the tournament outcome and the bracket scores is large even if the tournament outcome model specifies each probability correctly. The precision of the

estimated score of a single optimized bracket is low even if we assume that the accuracy is high.

In 2017, I presented research on multiple-entry strategies in standard scoring pools at the poster session of the New England Symposium on Statistics in Sports. The goal of this research was to test the quality of return on investment (ROI) estimates from a bracket strategy using a statistical hypothesis test.

A multiple-entry strategy was back-tested on 10 years of data from an office pool in Connecticut. The back-test involved predicting the best brackets to submit to these pools and then determining how these brackets would have performed if they had been submitted to these historical pools.

Historical pool data were available for the years 2008 through 2017. This was a standard scoring pool using exponential scoring (1, 2, 4, 8, 16, 32). The number of brackets submitted ranged from 178 to 241. The entry fee was $5. The pool awarded four prizes: 40%, 30%, 20%, and 10% of the pot. In the actual pools, ties were broken based on the best guess of the total score in the championship game—a guess had to be submitted with each bracket entry—but in this ROI analysis, the winnings were assumed to be divided in the case of a tie. The betting pattern in these pools had the characteristics noted in earlier research. The 1 seeds were over-bet. The hometown effect was evident: There was a home-state bias in favor of the Connecticut Huskies.

Calculating Bracket ROIs

Bracket ROIs were estimated using pool simulations. These used the probability model of a bracket pool described by Clair and Letscher. The inputs to modeling a pool are (1) N, the number of opponent entries, (2) a tournament outcome model, and (3) an opponent model. The tournament outcome is a random variable based on the tournament outcome model. Opponent brackets are random variables based on the opponent model. One pool simulation consists of one tournament outcome and N opponent brackets. The tournament outcome model was based on the Sagarin Predictor ratings. The opponent model was derived from either the Yahoo bracket contest pick distribution or the ESPN bracket contest pick distribution, using the heuristic from Clair and Letscher's research paper to estimate the head-to-head pick probabilities for each game. No adjustments to the pick distributions were made to address the home-state bias.

…this heuristic for deriving an opponent model is somewhat biased. Later review of the results indicated some bias: Simulated opponents were somewhat less likely to advance the most-popular champ pick than they should have been according to the pick advancement table for the nationwide bracket contest. A reanalysis performed using the mRchmadness method for deriving the opponent model indicated a slightly higher mean and a slightly higher variance in the ROIs, but overall, the results were not greatly improved by changing the opponent model.

The number of opponent entries used in these simulations was 200. In actual practice, the specific number of pool entries is not known in advance, so the number has to be estimated. Since pre-2008 pools had approximately 200 entries, 200 was a reasonable estimate. Bracket optimization results are relatively insensitive to errors in the estimated number of pool entries.

All the inputs used in this bracket pool probability model are available before tournament tipoff during the period while it is still possible to submit brackets to a pool. The Sagarin ratings for the season are available on Selection Sunday. The Yahoo and ESPN pick distributions are available and updated multiple times before the Thursday tournament tipoff. The number of opponent entries is estimated from past pools.

The bracket ROIs were based on 10,000 pool simulations for each year. For each simulated pool, the tournament outcome and all opponent scores that were among the top four scores were saved. This is the minimum information needed to determine the ROI of any additional brackets that may be added to the pool. This information allows us to estimate the ROI of any candidate bracket that we are considering for submission to the pool.

The ROI of a candidate bracket in a simulated pool is calculated by computationally "submitting" it to the pool and determining how it would have fared in the pool competition. The bracket is scored based on the tournament outcome for that pool. If the bracket score is not among the top four scores in the pool, then its ROI is minus $5 (minus one betting unit). That is, you just lost your entry fee.

If the bracket score is above the highest opponent score, then the bracket ROI will be 40% of the pot minus the entry fee, or 201 × 0.4 - 1 = 79.4 betting units or $397. If the score falls among the top four opponent scores, then its ROI depends on the bracket score rank and whether it ties one or more scores. The estimated average ROI of a bracket is the bracket's average ROI over the 10,000 simulated pools.

The estimated average ROI or expected ROI of a set of candidate brackets submitted to the pool can be calculated in a

similar manner. If a set of 10 brackets is submitted and none are in the money, then the ROI is –10 betting units. If one has the highest score and none of the others are in the top four scores, then the ROI is 210 × .4 - 10 = 74 betting units. The estimated expected ROI of the set is the average ROI for the 10,000 simulated pools.

Optimizing ROI

Clair and Letscher optimized the ROI of a single candidate bracket using a hill-climbing algorithm. It is necessary to generalize this process to mutually optimize multiple brackets.

The multiple-entry strategy uses an iterative algorithm. First, a single bracket is optimized using a hill-climbing algorithm. A commitment is made to submit this optimized bracket to the pool. Then the next bracket is optimized using the hill-climbing algorithm—but this next bracket is optimized relative to the pool consisting of the opponent brackets plus our previously submitted optimized bracket(s). That is, each additional bracket is optimized using knowledge of our previous submittals.

The hill-climbing algorithm was used to optimize picks for only the last seven games of the tournament. These seven games are the four regional championship games, the two semi-final games, and the final championship game. The optimization was limited to the last seven games of the tournament to reduce the computation time for what is already a computation-intensive algorithm.

For each candidate bracket, the hill-climb always starts with the same bracket. This is the bracket where the higher seed is picked to win each game. Then, starting with the championship game, different teams are tried as picks for that game and the bracket's expected ROI is calculated for each team. The bracket ROI is maximized over a range of possible picks for that game. To reduce computation time, some of the lower-seeded teams are not considered, since these lower-seeded teams are very unlikely to advance deep into the tournament. After the champ pick is optimized, the runner-up is optimized in the same manner. Then the two other regional champs are optimized.

Finally, the process is repeated again, starting with the champ. The process is repeated until repetition provides no improvement in the ROI.

This hill-climbing algorithm is different from the one that Clair and Letscher employed ... The Clair and Letscher algorithm was not tested in this application, so it is unclear whether their algorithm would be more accurate or less computation-intensive.

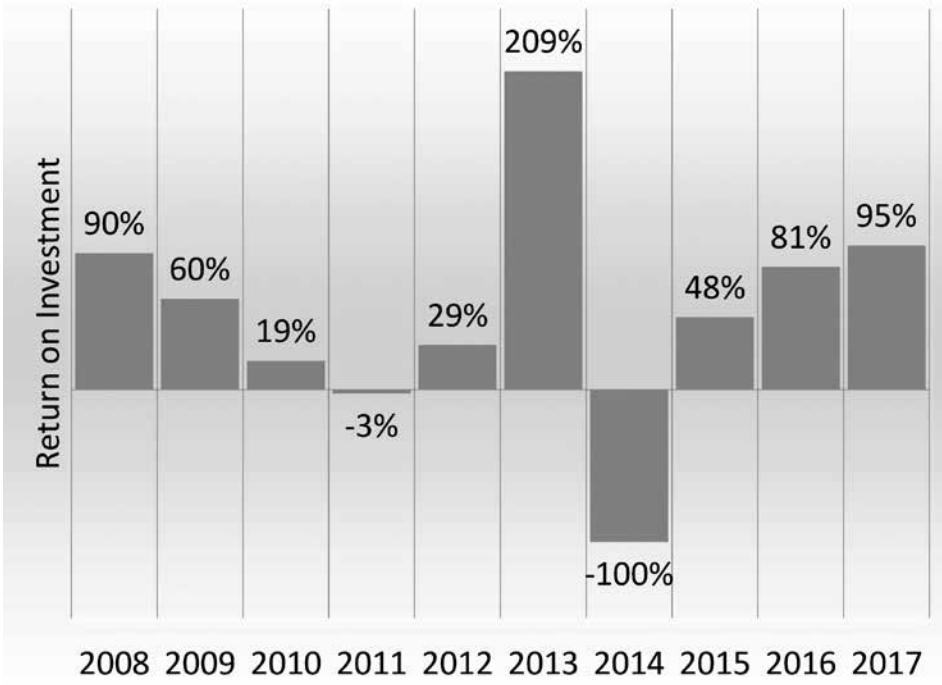


Figure 1. The ROI of 100 mutually optimized pool entries in a pool with approximately 200 opponent entries.

Results

Figure 1 shows the results of entering 100 mutually optimized entries in the 10 bracket pools. The ROIs ranged from 209% to –100%. The whole pot was won in the best year and the pool had more than 200 entries, resulting in an ROI of 209%. In the worst year, there were no winnings, leading to a 100% loss of all entry fees.

The two years when the multiple-entry strategy failed to show a profit were those when the Connecticut Huskies won the tournament. Since the office was in Connecticut, there was a large home-state bias (see Figure 2) in the pick distribution. The home-state bias is similar to estimates by Brad Null. Even a multiple-entry strategy that involved betting approximately 1/3 of the pool entries could not overcome this bias when the home-state team won the tournament.

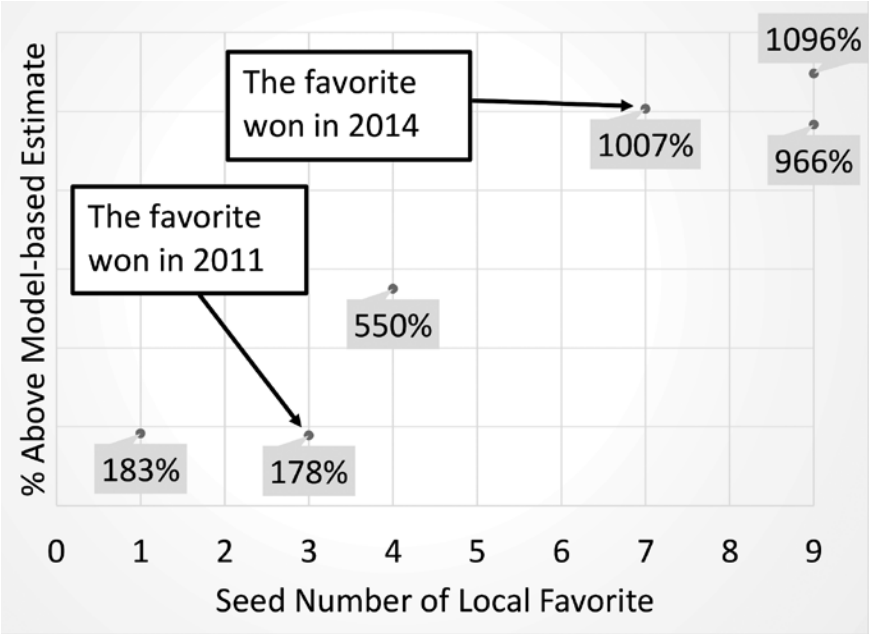


Figure 2. Home-state bias: the percentage by which the champ pick frequency of the local favorite exceeded its estimated value. The estimate was based on the pick advancement table from a nationwide bracket contest.

Adjusting the opponent model for home-state bias using Null's estimates is probably a good idea in general, but it would not have prevented these losses. Nor would such an adjustment help the strategy in any year when the home-state team wins because the more over-bet teams tend to be advanced less in the optimized brackets. Multiple-entry strategy for a single pool does not mitigate the risk of the home-state team winning the tournament.

Over the 10 years of the back-test, the multiple-entry strategy had an average ROI of 53%. The standard deviation of returns was 79%. The p-value of a one-tailed T-test is 3.2%, indicating grounds for accepting the hypothesis that this large-scale multiple-entry strategy is superior to the strategy employed by the average player.

A Practical Multiple-entry Strategy

Your pool czar will probably not let you bet one-third of the brackets in your pool, but you can generate a few brackets for multiple entries by varying your champ pick. Pick from among the strongest teams according to a predictive rating system…

Varying your champ pick is a good way to generate multiple entries for a bracket pool.

Further Reading

Adams, T. 2017. Modeling Multiple Entry Strategies in the NCAA Tournament Bracket Pool. Poster presentation at New England Symposium on Statistics in Sports. *http://nessis.org/nessis17/Adams.pdf*. Accessed May 31, 2017.

Adams, T. 2019. *Improving Your NCAA Bracket with Statistics*. Boca Raton: CRC Press.

Breiter, D.J., and Carlin, B.P. 1997. How to Play Office Pools if You Must. *CHANCE 10*:5–11.

Clair, B., and Letscher, D. 2007. Optimal Strategies for Sports Betting Pools. *Oper Res* 55:1,163–1,177.

ESPN. 2018. Who Picked Whom. *https://bit.ly/2FELGXd*. Accessed April 13, 2018.

Null, B. 2016. Homer Bias is real and it will derail your March Madness bracket. *https://bit.ly/2HlmjPQ*. Accessed March 3, 2018.

Shayer, E., and Powers, S. 2017. mRchmadness: Numerical Tools for Filling Out an NCAA Basketball Tournament Bracket. *https://bit.ly/2SVYHCo*. Accessed March 5, 2019.

Yahoo. 2018. Pick Distribution. *https://tournament.fantasysports.yahoo.com/t1/pickdistribution*. Accessed April 13, 2018.

About the Author

Tom Adams is a systems analyst who has spent most of his career in scientific research support and the creator of Poologic, a website that has provided research-based advice on winning bracket pools since 2000. Poologic has been featured in the *New York Times*, *Sports Illustrated*, *SmartMoney*, and other publications. Adams has published in statistics and probability. He has a BSc in mathematics from the University of North Carolina.

Nicole Lazar
Column Editor

Big Data and Privacy

When thinking about the topic of this issue's column, I batted various ideas around in my mind before settling on the theme of privacy in the age of Big Data. I knew I was on to something when, a few days later, my copy of *Significance* appeared in my mailbox and on the cover was the same theme!

I actually think about the questions of privacy and anonymity a fair bit, because I am by nature a private person, and don't like to share my information—online or otherwise—unnecessarily. At the same time, I do have an online presence as an academic. Going "off the grid" does not strike me as realistic or desirable.

Where is the balance? I'll admit that at times, I might go too far in the direction of discretion. An anecdote: When I was in my first assistant professor job, living in Pittsburgh, a local grocery chain had a loyalty card. When you bought groceries, you would swipe the card, and in return, the store would print coupons that matched

This article originally appeared in *CHANCE* 30.4.

your shopping patterns and would give you discounts from time to time, etc. The usual things.

One of my senior colleagues told me (with a gleam in his eye that only a true lover of data would have at such a moment) that the grocery chain had approached him about analyzing the masses of data that would be accumulating each day.

As a statistician and lover of data myself, I could see why he was intrigued. This was a rather early example of data mining: looking for patterns in shoppers' baskets of groceries. What's the likelihood that a customer will buy bread, milk, and eggs on the same trip? Do shoppers who buy beer tend to buy chips as well? You can think of your own examples. Again, this is all pretty common now, but at the time, for me at least, it seemed to open up new vistas of data analysis—and it was very exciting.

That's how the statistician in me reacted. The other me, the individual, didn't want one of those cards. I didn't want the grocery chain to know what I bought, when I bought it, what my patterns were. It was none of their business! When my boyfriend, now husband, left Pittsburgh, he gave me his loyalty card. And I'll admit, my first thought was, "Oh, good, now I can mess with their data!" I was amused to think about my colleague or one of our graduate students mining this huge accumulation of data and seeing a blip or switch in the record for that card.

Of course, I knew that it didn't really work that way, which brings me around to the bigger point—namely, how much anonymity or privacy do we really have in a situation like this?

The question was very much in the air as I write this column. A few days earlier, a story broke about Harvard University photographing classrooms in an effort to monitor student attendance. Cameras in lecture halls took pictures every minute, and the numbers of empty and full seats were counted, presumably by some pattern-recognition software. Harvard's Institutional Review Board ruled that the practice didn't fall under the purview of human-subject research. Faculty members were apparently informed about the study while students were kept in the dark. Had they known, students likely would have modified their behavior by coming to class more than usual. (Or, for those who like to mess with the data, maybe skipping class when they wouldn't ordinarily!)

The study was conducted in spring 2014 and made public in the fall. Students were outraged, as were many faculty, claiming that Harvard had invaded their privacy and administrators should have been more forthcoming about the project. It didn't help that this followed another scandal at Harvard, in which the administration had apparently been scanning emails of particular individuals.

Although school officials destroyed the classroom pictures immediately, and were at pains to insist that individual students per se were not of interest, many students, faculty members, and observers condemned the practice, blasting school officials on comment boards at the *Chronicle of Higher Education* and the *Harvard Crimson*.

Others, though, wondered what the fuss was all about, pointing out that data (including video and photographic) are collected about us all the time, so what's one more data mining experiment? Some also argued that there is no reasonable expectation to privacy in a college classroom, especially at a private institution that can, in many respects, set its own rules.

That this is one of the pertinent issues of the Big Data age is also made evident by a recent report—"Big Data and Privacy: A Technological Perspective"—submitted in May 2014 to President Obama by the President's Council of Advisors on Science and Technology (PCAST). Concerns about privacy and anonymity have long been coupled with changes in technology, dating back to the establishment of the U.S. Postal System in 1775, and on through the inventions of the telegraph, telephone, and portable cameras. The PCAST report surveyed the ways in which data collection, analysis, and use have converged in the modern age to make these issues particularly fraught.

In the past, it was easier—even feasible—for an individual to control what personal information was revealed in the public sphere. This is no longer the case. Public surveillance cameras and sensors record data without the individual being filmed or recorded necessarily even aware that it's happening. Additionally, social media such as Facebook and Twitter expose a great deal of personal information—sometimes intentionally, sometimes not.

The emergence of statistical techniques for the analysis of disparate data sources—a hallmark of Big Data—means that even if the information disclosed in one database maintains the individual's privacy, when taken together with other (also privacy-protecting) databases, identification of the individual is possible, and maybe inevitable in some cases. There are legal precedents governing some of these practices and concerns, but not all, and in any event, they'll all have to be revisited in light of the evolving technological landscape.

Part of the challenge inherent in the modern paradigm, as noted in the PCAST report, is that a certain

application may bring both benefits and harm, whether intended or not. For example, government is limited by the Fourth Amendment to the U.S. Constitution from searching private records in a home without probable cause. This seems straightforward until you think about what constitutes the boundaries of your home in a Wi-Fi, cloud-enabled world. The same technology that allows you to put family photographs and documents into, say, the cloud, to share with friends and relatives who live far away from you, also blurs the definition of what constitutes your home. Do the same legal protections extend there? If not, how might your privacy be compromised?

That may be a question for lawyers. How about one for statisticians? The PCAST report emphasizes that much of the benefit of Big Data comes from the data mining aspect—the ability to detect correlations that may be of interest or use. But it's important to keep in mind that these are just correlations, which means that the discovered relationships do not hold in all generality, and may not hold in particular for certain—and possibly vulnerable—sub-populations. Harm in a medical context, for example, could arise from mistaking correlations for hard-and-fast rules. And, as I've already noted, additional harm can come from merging data sets that were not meant to be merged, and subsequent analysis revealing facts about an individual that the person never wanted to make public.

The PCAST report also distinguishes between two types of data especially relevant for the privacy discussion: "born digital" and "born analog." Data that are born digital, as the name implies, are created specifically for use by a data-processing system. Examples include email and short message services, data entered into a computer or cell phone, location data from a GPS, "cookies" that track visits to websites, and metadata from phone calls. In all of these cases, there is intent, at some level, to provide the data to the monitoring system.

The challenge with data fusion is to devise analysis approaches that preserve the rights of individuals to control what they reveal about themselves to the world at large. This is an area of current research in statistics.

Privacy concerns here stem from two main sources: over-collection of data and fusion of data sources. Data over-collection occurs when the system, intentionally or otherwise, collects more information than what the user was aware of. An example from the PCAST report is an app called Brightest Flashlight Free, which millions of people have downloaded. Every time it was used, the app sent details of its location to the vendor. Why is this information necessary for a flashlight? Clearly, it isn't, and hence, this is an instance of over-collection.

The violation of privacy is obvious in that people who download a flashlight app for their phones are not expecting to reveal data about their locations every time they use it. To make it worse, the location information was apparently also sold to advertisers.

Data fusion is the term used when data from various sources are combined and analyzed together using data mining or other Big Data techniques. Even if each source on its own provides adequate privacy protection, they may draw a picture when multiple sources are analyzed together that is detailed enough to reveal specific and confidential information at the individual level. The challenge with data fusion is to devise analysis approaches that preserve the rights of individuals to control what they reveal about themselves to the world at large. This is an area of current research in statistics.

In contrast to data that are born digital, born analog data originate in the physical world and are created when some features of the physical world are detected by a sensor of some sort and converted to digital form. Examples include health statistics as collected by a Fitbit, imaging infrared video, cameras in video games that interpret hand or other gestures by the player, and microphones in meeting rooms.

When a camera takes pictures of a busy downtown neighborhood, it is bound to pick up some signal that is not of immediate interest. Hence born analog data will frequently contain more information than originally intended. Again, this can result in benefit as well as in harm, depending on how the data are used. Once

the born analog data are digitized, they can be fused with born digital data, just like any other source, and analyzed together as well.

Given this new state of reality, is it even possible to protect one's privacy? The answer to me seems to be a cautious yes. There is something of an arms race in which people work to crack protections, which in turn spurs the development of more sophisticated protective measures, which then pose a challenge to the first group, and on and on. Encryption is one way to make data more secure, although codes can be broken or stolen. "Notice and consent" is most often used for the protection of privacy in commercial settings. We all have encountered these when we want to install new software or a new app and are supposed to read and agree to a long list of terms before proceeding with the download. Among those terms are items relating to how your data may be used—for example, sold to advertisers or other third parties. But how often do you read those notices through and through? Always? Sometimes? Never? For most people, it's probably the last option, and this is a problem because it shifts the responsibility from the entity collecting your data back to you. The practice is even more problematic because not everyone is a lawyer. Even if we take the time to read the terms, many of us don't understand the implications and can't object if some of the terms seem unreasonable. In addition, the provider can change the privacy terms down the line without informing you of that fact.

Since notice and consent is the most prevalent model, and given the problems I just described, the PCAST report recommends that major effort be put into devising more workable and effective protections at this level, in part by placing the responsibility for using your data according to your wishes back with the providers—those collecting the data. One proposed framework is to have a variety of privacy protocols that emphasize different utilities. You would sign on to such a protocol, which would be passed on to app stores and the like when you use their services. The protocol you chose would dictate how your data could be used, shared, disseminated, etc. The groups offering the protocols could vet new programs and apps to ensure that they meet the desired standards. In either case, consumers would no longer have to wade through screens of legalese (or ignore them altogether!); the burden would be on the other parties to the transaction.

There is another perspective I should add, and that is symbolic data analysis, about which I have written here in the past. Recently my colleague Lynne Billard gave a talk in our weekly colloquium series about this approach to data analysis. She mentioned that one of the ways in which a statistician may receive a data set for which symbolic methods are appropriate is through aggregation. Her motivating example was from a medical data set, where the insurance company almost certainly doesn't care about my visits to the doctor; it cares about visits of "people like me." This provides a type of built-in anonymity for these massive, automatically generated data sets if they are indeed analyzed in an aggregated fashion.

Of course, this raises some other interesting and pertinent statistical questions: What does it mean to be "like me" for the purposes of this type of analysis? Who defines these categories? How sensitive are the analysis and its results to the particular aggregation scheme? Presumably the questions that the insurance company is asking should guide the aggregation. In some cases, maybe my age is relevant; in others, it may be my height, weight, cholesterol levels blood pressure, and so on. I can think of many dimensions in which someone may be "like me." Some will be important for certain types of analysis and not for others, but in any case, "I as me" will perhaps not be so critical.

These questions of data collection, analysis, and usage, and how they intersect with personal rights to (or desire for) anonymity and privacy, are not going to disappear. They are part of our cultural and technological landscape and are likely to expand over time. I think it's important for us to think about these issues and to decide where our personal lines are, how much we are comfortable sharing about ourselves, and what sort of presence we want to have in the digital and other modern realms.

To all of my students, former students, collaborators past and present, and old friends who try to connect via LinkedIn, Facebook, ResearchGate, and the like—when I don't respond, just know that I don't participate in any of those fora. It's my small way of keeping a corner of privacy in the world.

About the Author

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

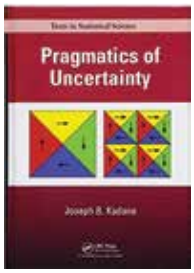
Further Reading

President's Council of Advisors on Science and Technology, May 2014. "Big Data and Privacy: A Technological Perspective."

Christian Robert
Column Editor

Pragmatics of Uncertainty

Joseph Kadane



Hardcover: 351 pages

Year: 2017

Publisher: Chapman and Hall/CRC Press

ISBN-13: 978-1498719841

I went through this book during the flight to the 2017 O'Bayes conference in Austin, somewhat paradoxically, since its message remains adamantly and unapologetically subjectivist. *Pragmatics of Uncertainty* is to be seen as a practical illustration of the *Principles of Uncertainty* Jay wrote in 2011 (and which I reviewed for *CHANCE*). Its avowed purpose is to allow the reader to check through Jay's applied work, whether or not he had "made good" on clearly setting out the motivations for his subjective Bayesian modeling. (While I presume the use of the same sequence "P of U" in both books is mostly a coincidence, I started wondering what a third "P of U" volume could be called. Perils of Uncertainty? Peddlers of Uncertainty? The game is afoot!)

The structure of the book is a collection of 15 case studies undertaken by the author over the past 30 years of his career, covering paleontology, survey sampling, legal knowledge, physics, climate, and even medieval Norwegian history. Each chapter starts with a short introduction that often explains how he came by the problem (most often, as an interesting PhD student consulting project at CMU), the difficulties in the analysis, and what became of his co-authors. As noted by the author, the main bulk of each chapter is the reprint (in a unified style) of the paper, and most of these papers are actually and freely available online. The chapter always concludes with an epilogue (or post-mortem) that reconsiders (very briefly) what had been done, what could have been done, and whether the Bayesian perspective was useful for the problem (unsurprisingly so for the majority of the chapters!). There are also reading suggestions in the other P of U and a few exercises. It thus comes as a complement for the classroom, if needed.

> *The purpose of the book is philosophical, to address, with specific examples, the question of whether Bayesian statistics is ready for prime time. Can it be used in a variety of applied settings to address real applied problems?*

The book is also a logical complement of the principles, helping to demonstrate how Jay himself applied his Bayesian principles to specific cases and how one can engage into the construction of a prior, a loss function, or a statistical model in identifiable parts that can then be criticized or reanalyzed.

I find browsing through this series of 15 problems fascinating and exhilarating, while I admire the dedication of Jay to every case he presents in the book. I also feel that this comes as a perfect complement to the earlier P of U, in that it makes referring to a complete application of a given principle most straightforward, with the problem being entirely described, analyzed, and—in most cases—solved within a given chapter.

A few chapters have discussions, originally published in the Valencia meeting proceedings or in another journal with discussions.

> *We think however that Bayes factors are overemphasized. In the very special case in which there are only two possible "states of the world," Bayes factors are sufficient. However in the typical case in which there are many possible states of the world, Bayes factors are sufficient only when the decision-maker's loss has only two values.* (p. 278) (Jay's reply to a discussion by John Skilling where he regrets the absence of marginal likelihoods in the chapter, a reply with which I completely subscribe.)

While all papers have been reset in the book style and fonts, and hence are seamlessly integrated, I do wish the original graphs had been edited as well, since they do not always look pretty. Although it would have implied a massive effort, it also would have been great had each chapter and problem been re-analyzed or at least discussed by a fellow Bayesian to illustrate the impact of individual modeling sensibilities. This may, however, be a future project for a graduate class, assuming all data sets are available, which is unclear from the text.

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

10 Great Ideas About Chance

Persi Diaconis and Brian Skyrms

Hardcover: 272 pages

Year: 2018

Publisher: Princeton University Press

ISBN-13: 978-0691174167

This book is a well-articulated collection of essays on chance, written jointly by Persi Diaconis, an American mathematician of Greek descent and former professional magician who is the Mary V. Sunseri Professor of Statistics and Mathematics at Stanford University, and Brian Skyrms, an American professor of logic and philosophy of science. As a warning, let me point out that I was a reviewer of this book for PUP, which means this review was adapted from the report I sent to the publisher.

The historical introduction ("measurement") of this book is most interesting, especially its analogy of chance with length. I would have appreciated a connection earlier than Cardano, such as some of the Greek philosophers, even though I gladly discovered that Cardano was not responsible only for the closed form solutions to the third-degree equation. I would also have liked to see more comments on the vexing issue of equiprobability: We all spend (if not waste) hours in the classroom explaining to (or arguing with) students why their solution is not correct, and they sometimes never get it! (We sometimes get it wrong as well.)

Why is such a simple concept so hard to explain or express? In short, although this is nothing but a personal choice, I would have made the chapter more conceptual and less chronologically historical.

> *Coherence is again a question of consistent evaluations of a betting arrangement that can be implemented in alternative ways.* (p. 46)

The second chapter, about Frank Ramsey, is definitely interesting, if only because it puts this "man of genius" back under the spotlight when he has been all but forgotten (at least in my circles) for joining probability and utility together, and for postulating that probability can be derived from expectations rather than the opposite. This is even though betting or gambling has a stigma in many cultures; at least, gambling for money, since most of our actions involve some degree of betting, but not in a rational or reasoned manner. (Of course, this is not a mathematical, but rather, a psychological objection.) Further, the justification through betting is somewhat tautological in that it assumes probabilities are true probabilities from the start. For instance, the Dutch book example on p. 39 produces a gain of .2 only if the probabilities are correct.

> *The force of accumulating evidence made it less and less plausible to hold that subjective probability is, in general, approximate psychology.* (p. 55)

A chapter on "psychology" may come as a surprise, but I feel a posteriori that it is appropriate. Most of it is about the Allais paradox, with entries on Ellesberg's distinction between risk and uncertainty, with only the former being quantifiable by "objective" probabilities, and on Tversky's and Kahneman's distinction between heuristics and the framing effect; i.e., how propositions are expressed affects the choice of decision-makers.

> *This is Bernoulli's swindle. Try to make it precise and it falls apart. The conditional probabilities go in different directions, the desired intervals are of different quantities, and the desired probabilities are different probabilities.* (p. 66)

The next chapter ("frequency") is about Bernoulli's Law of Large numbers and the stabilization of frequencies, with von Mises making it the basis of his approach to probability, and Birkhoff 's extension, which is capital for the development of stochastic processes and later for MCMC. I like the notions of "disreputable twin" (p. 63) and "Bernoulli's swindle," about the idea that "chance is frequency."

The authors call the identification of probabilities as limits of frequencies Bernoulli's swindle, because it cannot handle zero probability events. They offer a nice link with the testing fallacy of equating rejection of the null with acceptance of the alternative and an interesting description of how Venn perceived the fallacy but could not overcome it: "If Venn's theory appears to be full of holes, it is to his credit that he saw them himself."

The description of von Mises's Kollectiven (and the welcome intervention of Abraham Wald) clarifies my previous and partial understanding of the notion, although I am unsure it is that clear for all readers.

I also appreciate the connection with the very notion of randomness, which has not yet found, I fear, a satisfactory definition. This chapter asks more (interesting) questions than it answers (to those or others). But enough, this is a brilliant chapter!

> *…a random variable, the notion that Kac found mysterious in early expositions of probability theory.* (p. 87)

Chapter 5 ("mathematics") is very important, from my perspective, in that it justifies the necessity to associate measure theory with probability if one wishes to explore further than urns and dices, or to entitle Kolmogorov to posit his axioms of probability and to properly define conditional probabilities as random variables (as some of my students fail to realize). I very much enjoyed reading this chapter, but it may prove difficult for readers with no or little background in measure theory (some advanced mathematical details have vanished from the published version).

Still, this chapter constitutes a strong argument for preserving measure theory courses in graduate programs. As an aside, I find it amazing that mathematicians (including Kac!) had not at first realized the connection between measure theory and probability (p. 84), but it may not be so amazing given the difficulty many still have with the notion of conditional probability. (I would have liked to see some description of Borel's paradox when it is mentioned; p. 89).

> *Nothing hangs on a flat prior (…) Nothing hangs on a unique quantification of ignorance.* (p. 115)

The following chapter ("inverse inference") is about Thomas Bayes and his posthumous theorem, with an introduction setting the theorem at the center of the Hume-Price-Bayes triangle. (It is nice that the authors include a picture of the original version of the essay, as the initial title is much more explicit than the published version, as uncovered by Stephen Stiegler.) This is a short coverage, in tune with the fact that Bayes only contributed a 20-plus paper to the field, and it is logically followed by a second part (formerly another chapter) about Pierre-Simon Laplace.

Both parts focus on the selection of prior distributions on the probability of a binomial (coin tossing) distribution and emerging into a discussion of the position of statistics within or even outside mathematics. (The assertion that Fisher was the "Einstein of Statistics" on p. 120 may be disputed by many readers.)

> *So it is perfectly legitimate to use Bayes' mathematics even if we believe that chance does not exist.* (p. 124)

The seventh chapter is about Bruno de Finetti and his astounding representation of exchangeable sequences as mixtures of Independent and Identically Distributed (iid) sequences, defining an implicit prior on the side. While the description sticks to binary events, it quickly gets more advanced with the notion of partial and Markov exchangeability, including the most interesting connection between those exchangeabilities and sufficiency. (I would, however, disagree with the statement that "Bayes was the father of parametric Bayesian analysis" [p. 133], since this is extrapolating too much from The Essay.)

This may prove nonsensical, but I would have welcomed an entry at the end of the chapter on cases where the exchangeability representation fails; for instance, those cases when there is no sufficiency structure to exploit in the model. A bonus to the chapter is a description of the Birkhoff ergodic theorem "as a generalization of de Finetti" (p. 134–136), plus half-a-dozen pages of appendices on more technical aspects of de Finetti's theorem.

> *We want random sequences to pass all tests of randomness, with tests being computationally implemented.* (p. 151)

The eighth chapter ("algorithmic randomness") comes (again!) as a pleasant surprise since it centers on the character of Per Martin-Löf, who is little known in statistics circles. (The chapter starts with a picture of him with the iconic Oberwolfach sculpture in the background.) Martin-Löf's work concentrates on the notion of randomness, in a mathematical rather than probabilistic sense, and on its algorithmic consequences. I very much like the section on random generators, which includes a mention of our old friend RANDU, the 16-planes random generator!

This chapter connects with Chapter 4, since von Mises also attempted to define a random sequence, to the extent that it feels slightly repetitive (for instance, Jean Ville is mentioned in rather similar terms in both chapters). Martin-Löf's central notion is computability, which (kindly) forces us to visit Turing's machine, and its role in the undecidability of some logical statements, along with Church's recursive functions (with a potential link not exploited here to the notion of probabilistic programming, where one language is actually named Church, after Alonzo Church.)

I do not see how Martin-Löf's test for randomness can be implemented on a real machine, because the whole test requires going through the entire sequence: Since this notion connects with von Mises's Kollektivs, I am missing the point. Then Kolmororov is brought back with his own notion of complexity (which is also Chaitin's and Solomonov's).

Overall, this is a pretty challenging chapter both because of the notions it introduces and because I do not find it is completely conclusive about the notion(s) of randomness. (A side remark about casino hustlers and their "exploitation" of weak random generators: I believe Jeff Rosenthal has a similar, if maybe simpler, story about Canadian lotteries in his book.)

> *Does quantum mechanics need a different notion of probability? We think not.* (p. 180)

The penultimate chapter is about Boltzmann and the notion of "physical chance" or statistical physics, through a story that involves Zermelo and Poincaré,

and Gibbs, Maxwell, and the Ehrenfests. The discussion focuses on the definition of probability in a thermodynamic setting, opposing time frequencies to space frequencies, which requires ergodicity and hence Birkhoff (no surprise; this is about ergodicity), as well as von Neumann. This reaches a point where conjectures in the theory are still open.

What I always (if, presumably, naïvely) find fascinating in this topic is that ergodicity operates without requiring randomness. Dynamical systems can enjoy ergodic theorem, while being completely deterministic.

This chapter also discusses quantum mechanics, the main tenet of which requires probability, which has to be defined, from a frequency or a subjective perspective. The Bernoulli shift brings us back to random generators.

The authors briefly mention the Einstein-Podolsky-Rosen paradox, which sounds more metaphysical than mathematical in my opinion, although they go into great detail to explain Bell's conclusion that quantum theory leads to a mathematical impossibility (but they lost me along the way)—except that we "are left with quantum probabilities" (p. 183). The chapter leaves me still uncertain about why statistical mechanics carries the label "statistical," since it does not seem to involve inference at all.

> *If you don't like calling these ignorance priors on the ground that they may be sharply peaked, call them non-dogmatic priors or skeptical priors, because these priors are quite in the spirit of ancient skepticism.* (p. 199)

The last chapter ("induction") brings us back to Hume and the 18th century, where somehow "everything"—including statistics—started, except that Hume's strong skepticism (or skepticism) makes induction seemingly impossible (a perspective with which I agree to some extent, if not to Keynes's extreme version, when considering for instance financial time series as stationary. This is also a reason why I do not see the criticisms in *The Black Swan* as completely pertinent: They savage normality while accepting stationarity.)

The chapter re-discusses Bayes's and Laplace's contributions to inference as well, challenging Hume's conclusion of the impossibility to finer analyses, even though the representation of ignorance is not unique (p. 199). The authors call again for de Finetti's representation theorem as bypassing the issue of whether or not there is such a thing as chance. And escaping inductive skepticism. (The section about Goodman's grue hypothesis is somewhat distracting, maybe because I have always found this argument to be quite artificial and based on a linguistic pun rather than a logical contradiction.)

The part about (Richard) Jeffrey is quite new to me, but ends up quite abruptly, similarly to the one about Popper and his exclusion of induction. In this chapter, I appreciated very much the section on skeptical priors and its analysis from a meta-probabilist perspective.

There is no conclusion to the book, but to end up with a chapter on induction seems quite appropriate. There is an appendix as a probability tutorial, mentioning Monte Carlo resolutions, plus notes on all chapters and a commented bibliography. Definitely recommended! C

Independent Random Sampling Methods

Luca Martino, David Luengo, and Joaquín Míguez

Hardcover: 280 pages

Year: 2018

Publisher: Springer Verlag

ISBN-13: 978-3319726335

When I received a dedicated copy of this book from the first author, I was not aware of its existence. The three authors all work at Madrid universities and I have read (and blogged about) several papers of theirs about (population) Monte Carlo simulation in the recent years.

The book is definitely a pedagogical coverage of most algorithms used to simulate independent samples from a given distribution, which, of course, recoups some of the techniques exposed with more details by Luc Devroye in his 1986 *Non-Uniform Random Variate Generation* bible. It includes a whole chapter on accept-reject methods, in particular with a section about Payne-Dagpunar's band rejection that I had not seen previously. There is another entire chapter on ratio-of-uniforms techniques, on which the three authors had proposed generalizations covered by the book, years before I attempted to go the same way, having completely forgotten reading their paper at the time…or the much-earlier 1991 paper by Jon Wakefield, Alan Gelfand, and Adrian Smith.

The book also covers the "vertical density representation," due to Troutt (1991), which consists of considering the distribution of the density $p(.)$ of a random variable X as a random variable, $p(X)$. I remember pondering about this alternative to the cdf

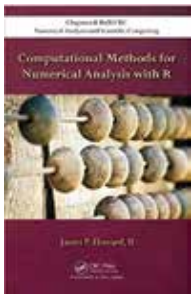
transform and giving up on it, since the outcome has a distribution depending on p, even when the density is monotonous. However, after reading the section and related papers, I am not certain that this is particularly appealing.

Given its title, the book obviously contains very little about MCMC. The exception is a last and final chapter that covers adaptive independent Metropolis-Hastings algorithms, in connection with some of the authors' recent work, such as multiple-try Metropolis. This relates to the (unidimensional) ARMS "ancestor" of adaptive MCMC methods.

All in all and with the bias induced by my working in the same area, I find the book quite a nice entry on the topic. It can be put to use directly in a Monte Carlo course at both undergraduate and graduate levels to avoid going into Markov chains. The book is certainly less likely to scare students away than the comprehensive *Non-Uniform Random Variate Generation* reference by Devroye (or a certain book by yours truly with a pumpkin-orange cover). On the contrary, it may well induce some of them to pursue a research career in this domain.

Computational Methods for Numerical Analysis with R

James Howard



Hardcover: 277 pages

Year: 2017

Publisher: CRC Press

SBN-13: 978-1498723633

This book consists of a traditional introduction to numerical analysis with backup from R codes and packages. The early chapters set the scenery, from basics on R to notions of numerical errors, before moving to linear algebra, interpolation, optimization, integration, differentiation, and ODEs. It comes with a package cmna that reproduces algorithms and testing.

While I do not find much originality in the book, given its adherence to simple resolutions of these topics, I could nonetheless use it for an elementary course in our first-year classes, with maybe the exception of the linear algebra chapter, which I did not find very helpful.

> *…you can have a solution fast, cheap, or correct, provided you only pick two.* (p. 27)

The (minor) issue I have with the book—which a potential mathematically keen student could face as well—is that there is little in the way of justifying a particular approach to a given numerical problem (as opposed to others) and in characterizing the limitations and failures of the presented methods (although this happens from time to time, such as inr gradient descent, p. 191). (Basking in my Gallic "mal-être," I am prone to over-criticize methods during classes, to the [increased] despair of my students, but I also feel that avoiding over-rosy presentations is a good way to prevent later disappointments or even disasters.)

In the case of this book, finding [more] ways of detecting would-be disasters would have been nice and to the point. A side comment as a statistician is that mentioning time series inter- or extra-polation without a statistical model sounds close to anathema! And makes extrapolation a weapon without a cause.

> *…we know, a priori, exactly how long the [simulated annealing] process will take since it is a function of the temperature and the cooling rate.* (p. 199)

Unsurprisingly, the section on Monte Carlo integration is disappointing for a statistician/probabilistic numericist like me, since it fails to give a complete-enough picture of the methodology. All simulations seem to proceed there from a large-enough hypercube, and recommending the "fantastic" (p. 171) R function integrate as a default is scary, given the ability of the selected integration bounds to misled its users.

Similarly, I feel that the simulated annealing section does not provide enough of a cautionary tale about the highly sensitive impact of cooling rates and absolute temperatures. It is only through the raw output of the algorithm applied to the traveling salesman problem that the novice reader can perceive the impact of some of these factors. (The acceptance bound on the jump (6.9) is, incidentally, wrongly called a probability on p. 199, since it can take values larger than 1.)

About the Author

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

DENVER, COLORADO

JSM2019 KEY DATES

ATTEND

May 1, 2019
Registration and Housing Open

May 31, 2019
Early Registration Deadline

June 1–June 29, 2019
Regular Registration

June 30–August 1
Late Registration

July 3
Housing Deadline

July 27–August 1
2019 JOINT STATISTICAL MEETINGS
Denver, Colorado

PARTICIPATE

January 23–April 3, 2019
Meeting and Event Request Submissions Accepted

January 25
Government Agency Registration Extension Request Deadline

April 1–18, 2019
Abstract Editing Open

April 15, 2019
Late-Breaking Session Proposals Deadline

May 17, 2019
Draft Manuscript Deadline

ww2.amstat.org/meetings/jsm/2019

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

BIG DATA, BIG OPPORTUNITIES.

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**

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.