Statistics and Science

A Report of the London Workshop on the Future of the Statistical Sciences

STATISTICS2013.ORG

INTERNATIONAL YEAR OF STATISTICS

ORGANIZING COMMITTEE MEMBERS

David Madigan, Columbia University (chair)

Peter Bartlett, University of California, Berkeley

Peter Bühlmann, ETH Zurich

Raymond Carroll, Texas A&M University

Susan Murphy, University of Michigan

Gareth Roberts, University of Warwick

Marian Scott, University of Glasgow

Simon Távare, Cancer Research United Kingdom Cambridge Institute

Chris Triggs, University of Auckland

Jane-Ling Wang, University of California, Davis

Ronald L. Wasserstein, American Statistical Association (staff)

Khangelani Zuma, HSRC South Africa

ALPHABETICAL LIST OF SPONSORS OF THE 2014 LONDON WORKSHOP ON THE FUTURE OF THE STATISTICAL SCIENCES

American Statistical Association

Bernoulli Society

Institute of Mathematical Statistics

International Association for Statistical Education

International Association of Survey Statisticians

International Chinese Statistical Association

International Indian Statistical Association

International Society for Business and Industrial Statistics

International Statistical Institute

National Science Foundation

Royal Statistical Society

SAGE

Statistical Society of Canada

Taylor & Francis Group

Wellcome Trust

Wiley-Blackwell

STATISTICS2013.ORG
INTERNATIONAL YEAR OF
STATISTICS

CONTENTS

4 EXECUTIVE SUMMARY

6 INTRODUCTION

8 SECTION 1
How Statistics Is Used in the Modern World: Case Studies

20 SECTION 2
Current Trends and Future Challenges in Statistics: Big Data

26 SECTION 3
Current Trends and Future Challenges in Statistics: Other Topics

34 SECTION 4
Conclusion

37 ACKNOWLEDGEMENTS

EXECUTIVE SUMMARY

The American Statistical Association, the Royal Statistical Society, and four other leading statistical organizations partnered in celebrating 2013 as the International Year of Statistics. The capstone event for this year of celebration was the Future of the Statistical Sciences Workshop, held in London on November 11 and 12, 2013. This meeting brought together more than 100 invited participants for two days of lectures and discussions. As well as an invited audience who were present for the event, the organizers made the lectures are freely available to the public online at *www.statisticsviews.com* (registration required).

Statistics can be most succinctly described as the science of uncertainty. While the words "statistics" and "data" are often used interchangeably by the public, statistics actually goes far beyond the mere accumulation of data. The role of a statistician is:

- To design the acquisition of data in a way that minimizes bias and confounding factors and maximizes information content
- To verify the quality of the data after it is collected
- To analyze data in a way that produces insight or information to support decision-making

These processes always take into explicit account the stochastic uncertainties present in any real-world measuring process, as well as the systematic uncertainties that may be introduced by the experimental design. This recognition is an inherent characteristic of statistics, and this is why we describe it as the "science of uncertainty," rather than the "science of data."

Data are ubiquitous in 21st-century society: They pervade our science, our government, and our commerce. For this reason, statisticians can point to many ways in which their work has made a difference to the rest of the world. However, the very usefulness of statistics has worked in some ways as an obstacle to public recognition. Scientists and executives tend to think of statistics as infrastructure, and like other kinds of infrastructure, it does not get enough credit for the role it plays. Statisticians, with some prominent exceptions, also have been unwilling or unable to communicate to the rest of the world the value (and excitement) of their work.

This report, therefore, begins with something that was mostly absent from the London workshop: seven case studies of past "success stories" in statistics, which in all cases have continued to the present day. These success stories are certainly not exhaustive—many others could have been told—but it is hoped that they are at least representative. They include:

- The development of the randomized controlled trial methodology and appropriate methods for evaluating such trials, which are a required part of the drug development process in many countries.
- The application of "Bayesian statistics" to image processing, object recognition, speech recognition, and even mundane applications such as spell-checking.
- The explosive spread of "Markov chain Monte Carlo" methods, used in statistical physics, population modeling, and numerous other applications to simulate uncertainties that are not distributed according to one of the simple textbook models (such as the "bell-shaped curve").
- The involvement of statisticians in many high-profile court cases over the years. When a defendant is accused of a crime because of the extraordinary unlikelihood of some chain of events, it often falls to statisticians to determine whether these claims hold water.
- The discovery through statistical methods of "biomarkers"—genes that confer an increased or decreased risk of certain kinds of cancer.
- A method called "kriging" that enables scientists to interpolate a smooth distribution of some quantity of interest from sparse measurements. Application fields include mining, meteorology, agriculture, and astronomy.
- The rise in recent years of "analytics" in sports and politics. In some cases, the methods involved are not particularly novel, but what is new is the recognition by stakeholders (sports managers and politicians) of the value that objective statistical analysis can add to their data.

Undoubtedly the greatest challenge and opportunity that confronts today's statisticians is the rise of Big Data—databases on the human genome, the human brain, Internet commerce, or social networks (to name a few) that dwarf in size any databases statisticians encountered in the past. Big Data is a challenge for several reasons:

- *Problems of scale.* Many popular algorithms for statistical analysis do not scale up very well and run hopelessly slowly on terabyte-scale data sets. Statisticians either need to improve the algorithms or design new ones that trade off theoretical accuracy for speed.
- *Different kinds of data.* Big Data are not only big, they are complex and they come in different forms from what statisticians are used to, for instance images or networks.
- *The "look-everywhere effect."* As scientists move from a hypothesis-driven to a data-driven approach, the number of spurious findings (e.g., genes that appear to be connected to a disease but really aren't) is guaranteed to increase, unless specific precautions are taken.
- *Privacy and confidentiality.* This is probably the area of greatest public concern about Big Data, and statisticians cannot afford to ignore it. Data can be anonymized to protect personal information, but there is no such thing as perfect security.
- *Reinventing the wheel.* Some of the collectors of Big Data—notably, web companies—may not realize that statisticians have generations of experience at getting information out of data, as well as avoiding common fallacies. Some statisticians resent the new term "data science." Others feel we should accept the reality that "data science" is here and focus on ensuring that it includes training in statistics.

Big Data was not the only current trend discussed at the London meeting, and indeed there was a minority sentiment that it is an overhyped topic that will eventually fade. Other topics that were discussed include:

- *The reproducibility of scientific research.* Opinions vary widely on the extent of the problem, but many "discoveries" that make it into print are undoubtedly spurious. Several major scientific journals are requiring or encouraging authors to document their statistical methods in a way that would allow others to reproduce the analysis.
- *Updates to the randomized controlled trial.* The traditional RCT is expensive and lacks flexibility. "Adaptive designs" and "SMART trials" are two modifications that have given promising results, but work still needs to be done to convince clinicians that they can trust innovative methods in place of the tried-and-true RCT.
- *Statistics of climate change.* This is one area of science that is begging for more statisticians. Climate models do not explicitly incorporate uncertainty, so the uncertainty has to be simulated by running them repeatedly with slightly different conditions.
- *Statistics in other new venues.* For instance, one talk explained how new data capture methods and statistical analysis are improving (or will improve) our understanding of the public diet. Another participant described how the United Nations is experimenting for the first time with probabilistic, rather than deterministic, population projections.
- *Communication and visualization.* The Internet and multimedia give statisticians new opportunities to take their work directly to the public. Role models include Nate Silver, Andrew Gelman, Hans Rosling, and Mark Hansen (two of whom attended the workshop).
- *Education.* A multifaceted topic, this was discussed a great deal but without any real sense of consensus. Most participants at the meeting seemed to agree that the curriculum needs to be re-evaluated and perhaps updated to make graduates more competitive in the workplace. Opinions varied as to whether something needs to be sacrificed to make way for more computer science–type material, and if so, what should be sacrificed.
- *Professional rewards.* The promotion and tenure system needs scrutiny to ensure non-traditional contributions such as writing a widely used piece of statistical software are appropriately valued. The unofficial hierarchy of journals, in which theoretical journals are more prestigious than applied ones and statistical journals count for more than subject-matter journals, is also probably outmoded.

In sum, the view of statistics that emerged from the London workshop was one of a field that, after three centuries, is as healthy as it ever has been, with robust growth in student enrollment, abundant new sources of data, and challenging problems to solve over the next century. ❖

INTRODUCTION

In 2013, six professional societies declared an International Year of Statistics to celebrate the multifaceted role of statistics in contemporary society, to raise public awareness of statistics, and to promote thinking about the future of the discipline. The major sponsors of the yearlong celebration were the American Statistical Association, the Royal Statistical Society, the Bernoulli Society, the Institute of Mathematical Statistics, the International Biometric Society, and the International Statistical Institute. In addition to these six, more than 2,300 organizations from 128 countries participated in the International Year of Statistics.

The year 2013 was a very appropriate one for a celebration of statistics. It was the 300th anniversary of Jacob Bernoulli's *Ars conjectandi* (Art of Conjecturing) and the 250th anniversary of Thomas Bayes' "An Essay Towards Solving a Problem in the Doctrine of Chances." The first of these papers helped lay the groundwork for the theory of probability. The second, little noticed in its time, eventually spawned an alternative approach to probabilistic reasoning that has truly come to fruition in the computer age. In very different ways, Bernoulli and Bayes recognized that *uncertainty* is subject to mathematical rules and rational analysis. Nearly all research in science today requires the management and calculation of uncertainty, and for this reason statistics—the science of uncertainty—has become a crucial partner for modern science.

Statistics has, for example, contributed the idea of the randomized controlled trial, an experimental technique that is universal today in pharmaceutical and biomedical research and many other areas of science. Statistical methods underlie many applications of machine reasoning, such as facial recognition algorithms. Statistical analyses have been used in court on numerous occasions to assess whether a certain combination of events is incriminating or could be just a coincidence. New statistical methods have been developed to interpret data on the human genome and to detect biomarkers that might indicate a higher risk for certain kinds of cancer. Finally, sound statistical reasoning is transforming the sports that we play and the elections we vote in. All of these statistical "success stories," and more, are discussed in detail later in this report.

The International Year of Statistics came at a time when the subject of statistics itself stood at a crossroads. Some of its most impressive achievements in the 20th century had to do with extracting as much information as possible from relatively small amounts of data—for example, predicting an election based on a survey of a few thousand people, or evaluating a new medical treatment based on a trial with a few hundred patients.

While these types of applications will continue to be important, there is a new game in town. We live in the era of Big Data. Companies such as Google or Facebook gather enormous amounts of information about their users or subscribers. They constantly run experiments on, for example, how a page's layout affects the likelihood that a user will click on a particular advertisement. These experiments have millions, instead of hundreds, of participants, a scale that was previously inconceivable in social science research. In medicine, the Human Genome Project has given biologists access to an immense amount of information about a person's genetic makeup. Before Big Data, doctors had to base their treatments on a relatively coarse classification of their patients by age group, sex, symptoms, etc. Research studies treated individual variations within these large categories mostly as "noise." Now doctors have the prospect of being able to treat every patient uniquely, based on his or her DNA. Statistics and statisticians are required to put all these data on individual genomes to effective use.

The rise of Big Data has forced the field to confront a question of its own identity. The companies that work with Big Data are hiring people they call "data scientists." The exact meaning of this term is a matter of some debate; it seems like a hybrid of a computer scientist and a statistician. The creation of this new job category brings both opportunity and risk to the statistics community. The value that statisticians can bring to the enterprise is their ability to ask and to answer such questions as these: Are the data representative? What is the nature of the uncertainty? It may be an uphill battle even to convince the owners of Big Data that their data are subject to uncertainty and, more importantly, bias.

On the other hand, it is imperative for statisticians not to be such purists that they miss the important scientific developments of the 21st century. "Data science" will undoubtedly be somewhat different from the discipline that statisticians are used to. Perhaps statisticians will have to embrace a new identity. Alternatively, they might have to accept the idea of a more fragmented discipline in which standard practices and core knowledge differ from one branch to another.

These developments formed the background for the Future of the Statistical Sciences Workshop, which was held on November 11 and 12, 2013, at the offices of the Royal Statistical Society in London. More than 100 statisticians, hailing from locations from Singapore to Berkeley and South Africa to Norway, attended this invitation-only event, the capstone of the International Year of Statistics. The discussions from the workshop comprise the source material for Sections 2 and 3 of this document.

Unlike the workshop, this report is intended primarily for people who are not experts in statistics. We intend it as a resource for students who might be interested in studying statistics and would like to know something about the field and where it is going, for policymakers who would like to understand the value that statistics offers to society, and for people in the general public who would like to learn more about this often misunderstood field. To that end, we have provided in Section 1 some examples of the use of statistics in modern society. These examples are likely to be familiar to most statisticians, but may be unfamiliar to other readers.

One common misconception about statisticians is that they are mere data collectors, or "number crunchers." That is almost the opposite of the truth. Often, the people who come to a statistician for help—whether they be scientists, CEOs, or public servants—either can collect the data themselves or have already collected it. The mission of the statistician is to work with the scientists to ensure that the data will be collected using the optimal method (free from bias and confounding). Then the statistician extracts meaning from the data, so that the scientists can understand the results of their experiments and the CEOs and public servants can make well-informed decisions.

Another misperception, which is unfortunately all too common, is that the statistician is a person brought in to wave a magic wand and make the data say what the experimenter wants them to say. Statisticians provide researchers the tools to declare comparisons "statistically significant" or not, typically with the implicit understanding that statistically significant comparisons will be viewed as real and non-significant comparisons will be tossed aside. When applied in this way, statistics becomes a ritual to *avoid* thinking about uncertainty, which is again the opposite of its original purpose.

Ideally, statisticians should provide concepts and methods to learn about the world and help people make decisions in the face of uncertainty. If anything is certain about the future, it is that the world will continue to need this kind of "honest broker." It remains in question whether statisticians will be able to position themselves not as number crunchers or as practitioners of an arcane ritual, but as data explorers, data diagnosticians, data detectives, and ultimately as answer providers. ❖

SECTION 1. How Statistics Is Used in the Modern World: Case Studies

In this part of the report, we present seven case studies of the uses of statistics in the past and present. We do not intend these examples to be exhaustive. We intend them primarily as educational examples for readers who would like to know, "What is statistics good for?" Also, we intend these case studies to help frame the discussion in Sections 2 and 3 of current trends and future challenges in statistics.

1.1 Randomized Controlled Trials

Every new pharmaceutical product in the United States and many other countries goes through several rounds of statistical scrutiny before it can reach the marketplace. The prototypical type of study is called a randomized controlled trial, an experimental design that emerged from Sir Ronald Fisher's research nearly a century ago.

In 1919, the Cambridge-educated geneticist and statistician accepted a position at the Rothamsted Experimental Station, an agricultural research facility in Hertfordshire, England. While working there, he clarified many of scientists' previously haphazard ideas about experimental design, and his ideas had repercussions that went far beyond agronomy.

Here is a typical problem of the type Fisher analyzed: A researcher wants to know if a new fertilizer makes corn more productive. He could compare a sample of plants that have been given the fertilizer (the "treatment" group) with plants that have not (the "control" group). This is a controlled trial. But if the treatment group appeared more productive, a skeptic could argue that those plants had come from more vigorous seeds, or had been given better growing conditions.

To anticipate such objections, the treatment and control group should be made as similar to each other in every possible way. But how can one enforce this similarity? What is to keep the experimenter from inadvertently or deliberately stacking the deck? Fisher's answer was revolutionary and far from obvious: *randomization*. If the treatment (the fertilizer) is given to random plants in random plots, the experimenter cannot affect the results with his own bias.

Randomization seems counterintuitive at first, because there is no attempt to match the treatment group and control group. But in fact, it exploits the laws of probability. If you flip a coin 100 times, you are much more likely to get a roughly even split of heads and tails than you are to get all heads, or even 75 percent heads. Similarly, in a controlled experiment, randomness is a rough (though not exact) guarantee of fairness.

Besides eliminating bias and approximately matching the treatment and control groups, the randomized controlled trial (RCT) design has one more advantage. It makes the source of uncertainty explicit so that it can be modeled mathematically and used in the analysis. If the uncertainty lay in the quality of the seed or the soil, it would be difficult for an experimenter to model. But in an RCT, the randomization procedure itself is the source of uncertainty. In 100 flips of a coin, it's easy to say what is a reasonable and an unreasonable number of heads to expect. As a result, the researcher can *quantify* the uncertainty. When assessing whether the fertilizer works, he can calculate a statistical measure (a "p-value") that reflects the strength of the evidence that it does. (See sidebar, "Statisticians Were Here." Also see §1.4 for some downsides to the uncritical use of p-values.)

In medicine, the use of controlled trials, and even randomization, goes back a long way. James Lind's experiments on the treatment of scurvy in 1747, which showed that lemons and limes were the most effective treatment out of six alternatives, are often cited as the first controlled trials in history. In 1835, the pharmacists of Nuremberg, Germany, added the idea of randomization in an experiment to tell whether a homeopathic "remedy" could be distinguished from a placebo. However, these experiments did not yet have Fisher's mathematical framework to quantify the uncertainty.

It was only after World War II that Austin Bradford Hill, a British epidemiologist, conducted the first modern RCT in medicine. In 1948, he demonstrated overwhelming evidence that the newly discovered antibiotic vancomycin was effective against tuberculosis. His study was a watershed moment—for medicine and for statistics. One of the discoverers of vancomycin won the Nobel Prize in 1952. Tuberculosis, one of the greatest scourges of the 19th and early 20th centuries, suddenly became a manageable disease. And RCTs came into great demand, as the success of "wonder drugs" like penicillin and vancomycin made the development of new pharmaceuticals into a highly lucrative business.

Some non-statistical factors also contributed to the ascendance of RCTs in medical research. In 1937, more than 100 people died from a new "wonder drug," sulfanilamide, not because of the drug, but because of the solvent in which it was suspended. This tragedy motivated the Food, Drug, and Cosmetic Act, passed in 1938, which required drug manufacturers to provide evidence of safety to the Food and Drug Administration (FDA). In 1961, the public outcry over thalidomide (an experimental drug that was shown to cause severe birth defects) led to the passage of the Kefauver-Harris Drug Amendment, which required "adequate and well-controlled studies" to establish effectiveness and safety for new drugs for the first time. These studies are often, though not always, randomized.

If there is any problem with the RCT, it has been too successful, to the point of becoming a straitjacket. "It was one of the greatest inventions in medical history," says Don Berry of MD Anderson Cancer Center. "The only problem was that people didn't want to tinker with it."

During the AIDS epidemic of the 1980s, RCTs came under fire for being too slow and too insensitive to patients' needs. The agitation of AIDS activists led to certain reforms such as easier access to experimental drugs and the use of "surrogate endpoints" (such as improved T-cell counts) that could be used as evidence of effectiveness. These surrogates themselves had to be investigated statistically. Other innovations include the use of "historical controls," interim analysis (i.e., analysis of data while the study is still in progress) and early termination of studies in which the treatment has either an extremely positive or an extremely negative effect. Thus, the RCT does not have to be a straitjacket. However, the involvement of statisticians has become even more important to ensure that such modifications do not compromise the trial's scientific validity.

Another new challenge to traditional RCTs is personalized medicine. Doctors realize now that cancer is not one disease, but has many subtypes (see §5), each potentially requiring a different treatment. Thanks to genomics, doctors will increasingly be able

For many years, the field of statistics has had two philosophical camps with different answers to a fundamental question: What does "probability" mean? The camps are known as the "frequentists" and the "Bayesians."

to distinguish different types of patients, as well. The conventional RCT with hundreds or thousands of patients may become simply impossible. There may not even be a thousand patients in the world with cancer subtype A and genetic markers B, C, and D. Section 3 will discuss one new approach (adaptive designs) that will enable researchers to zero in on effective treatments for smaller populations. It remains to be seen whether such methods will achieve the level of acceptance that traditional RCTs have.

1.2 The Bayesian Paradigm and Image Processing

For many years, the field of statistics has had two philosophical camps with different answers to a fundamental question: What does "probability" mean? The camps are known as the "frequentists" and the "Bayesians." The debate is not merely academic, because different viewpoints on this question lead to different methodologies. However, in recent years, the controversy has diminished and statisticians have come to realize that both viewpoints can be useful in different contexts.

In brief, the frequentist viewpoint is that a probability reflects how often a particular outcome will be observed in repeated trials of the same experiment. The language of the frequentists pervades statistical textbooks; the examples, such as drawing balls from an urn or throwing dice, are ideal situations in which the same procedure can be repeated many times with uncertain results. The frequentist paradigm—developed by early pioneers such as Fisher, Jerzy Neyman, and Karl Pearson—is reflected in the classical design of clinical trials (§1), where the results are phrased in terms of what would happen if the experiment were repeated many times.

The Bayesian philosophy, named after the Reverend Thomas Bayes (see Introduction), applies the mathematics of probability more generally, not just to long-run frequencies, but also to the probabilities of unique events such as the "probability that candidate A will win the election." Often, the Bayesian view of probability is described as a "degree of belief" in a statement, but Andrew Gelman, a Bayesian statistician, has argued that this interpretation is in no way obligatory. A Bayesian statistician is free to interpret probability in whatever way best suits the problem—as a frequency, as a degree of belief, or simply as a function that obeys the mathematical rules of probability.

Some recent work in cognitive psychology has drawn interesting connections between frequentist and Bayesian ideas. From one direction, Josh Tenenbaum, Tom Griffiths, and others have had striking success modeling human inference and decisionmaking as being approximately Bayesian. From the other side, Gerd Gigerenzer has demonstrated that people understand uncertainty much better when framed as frequencies, rather than probabilities. Various probability problems become much less confusing for people when the probabilities are reframed as frequencies (e.g., 85 out of 100, rather than a probability of 85%).

Bayesian statistics takes its name from Bayes' theorem, which is a rule for updating our belief in a hypothesis as we collect new evidence. One version of it can be stated as follows:

Posterior odds = prior odds × likelihood ratio.

A good example of Bayes's rule is provided by spell-checking programs. Suppose, for instance, a user types the word "radom" and the computer has to decide whether she meant to type "random" or "Radom," the city in Poland. Consulting the Google language database, the computer determines that the word "random" appears 200 times as often as "Radom" in all documents. In the absence of any other information, the "prior odds" are 200:1 in favor of "random." However, a spell-check program that simply defaulted to the most common word all the time would change every word to "the." So the prior odds have to be modified by the evidence of what the typist actually typed. According to Google's model of spelling errors, it is 500 times more likely that typists will type "radom" if the word they meant to type is "Radom" (which they will do with probability 0.975) than if the word is "random" (probability 0.00195). So the likelihood ratio is 1/500, and the posterior odds become (200/1)(1/500), or 2:5. Thus the spell checker will not auto-correct the word. On the other hand, if the spell-checker knew that the word came from a document on statistics, the prior odds in favor of "random" would go up and the spell-checker would then auto-correct the word. Or if the typist were sloppy, the likelihood ratio of an error versus a correct spelling would go up, and again the posterior odds would shift in favor of "random." This shows how easily Bayes' rule incorporates new information.

Humans update their beliefs every time they look at something. "Perceptions are predictive, never entirely certain, hypotheses of what may be there," wrote Richard Gregory, an experimental psychologist and expert on visual illusions. Ordinarily, we assume that the source of illumination in a scene is at the top. We assume that solid objects are more likely to be convex than concave. These are prior beliefs, which visual illusions exploit to create images that confuse us. But most of the time, our hypotheses serve us well. Images are inherently ambiguous—they are projections of a three-dimensional world onto a two-dimensional retina—and hence we need assumptions to make sense of what we see. We constantly refine or discard them as we get new visual or sensory data. Our unconscious hypotheses allow us to separate foreground from background, to read a blurry sign in the distance, to recognize faces—all tasks that are quite difficult for a computer.

However, recent research has helped machines figure out more about the content of an image by using Bayesian reasoning. For example, many digital cameras have the ability to "lock onto" faces. They will draw a little rectangle around anything that the camera's "brain" thinks is likely to be a face. The technology involved is surprisingly recent—it was invented by Paul Viola and Michael Jones in 2001—yet it has become nearly ubiquitous.

For another example, the Microsoft Kinect game player uses Bayesian algorithms to track a user's motions. It is programmed to make certain assumptions about how images are generated, just as humans do: Scenes contain objects, objects have textures, textures reflect light in certain ways. These causal relationships constrain our prior hypotheses about a scene. They do the same thing when programmed into a computer. When a new image comes in, the software can filter it through this network of assumed relationships (called a "Bayesian network") and generate the most likely hypothesis about what is foreground and what is background, where your hands are, and which hand is connected to which shoulder.

Of course, this research raises the question of what kinds of prior hypotheses are imbedded in our own brains and how humans arrive at them. It's reasonable to expect that such questions will drive collaboration between statisticians and psychologists for a long time to come.

1.3 The Markov Chain Monte Carlo Revolution

Statistics was a multidisciplinary science from the very beginning, long before that concept became fashionable. The same techniques developed to analyze data in one application are very often applicable in numerous other situations. One of the best examples of this phenomenon in recent years is the application of Markov Chain Monte Carlo (MCMC) methods. While MCMC was initially invented by statistical physicists who were working on the hydrogen bomb, it has since been applied in settings as diverse as image analysis, political science, and digital humanities.

Markov Chain Monte Carlo is essentially a method for taking random samples from an unfathomably large and complex probability distribution. For a simple example, a prison official once brought statistician Persi Diaconis a message between two prisoners that had been intercepted. The message was written in a code that did not resemble the English alphabet and the guards had not been able to decipher it. Diaconis gave it to a student as a challenge. Remarkably, the student succeeded on the first try, using an MCMC algorithm.

Here's how it worked. The "large probability distribution" describes all possible ways that the alphabet could be encoded into 26 symbols. Not all ways are equally likely. If one proposed decryption produces a word with the letters "QA" adjacent to one another, this is a highly implausible decryption. On the other hand, a letter combination one expects to see often is "TH," so a decryption that produces a lot of these is quite plausible.

The algorithm takes a random walk through this space of all possible decryptions. It starts with a randomly chosen decryption. Then, at each step, it considers one possible revision. If the symbol for "A" is changed to "U," the "QAs" would become much more plausible "QUs." Each time it considers a change, the MCMC algorithm computes the plausibility score of the new decryption. If the new decryption is more plausible than the old one, the algorithm makes that change. If not, it will *probably*, but not necessarily, reject the change. Sometimes, it will accept a change that is *a priori* less likely. This keeps the algorithm from becoming "trapped" in a dead end. Like a human detective, it sometimes needs to try out alternative hypotheses that seem less plausible at first.

Eventually, after many iterations, MCMC will arrive at a random sample from the space of plausible decryptions. In the case of the prisoner's code there is *only one* plausible decryption, so the "random sample" *is* the solution. Diaconis and his student knew they had found it when, after a few thousand steps of the MCMC algorithm, the computer came up with the following decryption: "To bat-rb. Con todo mi respeto. I was sitting down playing chess with danny…"

While secret messages between prisoners are a rather unusual application of MCMC, the method has a mind-boggling array of other uses. The original algorithm was designed in the late 1940s by Nicholas Metropolis, Stanislaw Ulam, Edward Teller, and others to simulate the motion of neutrons in an imploding hydrogen bomb. This motion is essentially random. However, "random" does not mean "arbitrary." The neutrons obey physical laws, and this makes certain outcomes much more likely than others. The probability space of all plausible neutron paths is far too large to store in a computer, but Metropolis' algorithm enables the computer to pick random plausible paths and thereby predict how the bomb will behave.

In a completely different application, MCMC has been used to analyze models of how politicians vote on proposed legislation or how U.S. Supreme Court justices vote on cases that come before them. The second example is of particular interest because the justices typically say very little in public about their political viewpoints after their confirmation hearings, yet their ideologies can and do change quite a bit during the course of their careers. Their votes are the only indicator of these changes. While political pundits are always eager to "read the tea leaves," their analysis typically lacks objectivity and quantitative rigor.

By contrast, the statistical approach produces an explicit measure of where each justice lies along a liberal-conservative spectrum; the changes can be graphed over time to produce an immediately understandable picture of each justice's career. The graphs make it easy to spot the court's "median voter," who has historically been the swing vote in many closely divided cases. The swing vote for many years was Byron White; then it was Sandra Day O'Connor; and, in recent years, it has been Anthony Kennedy. Admittedly, this is not too surprising for Supreme Court buffs, but what's remarkable is that a computer equipped with this method can "figure it out" with no prior knowledge of politics or the legal details of the cases involved.

In this example, MCMC is used to support an explicitly Bayesian analysis. One technical problem with Bayes' rule is that the likelihood ratio (discussed in §1.2) is often very difficult to calculate, because it involves summing (or in mathematical jargon, "integrating") over an inordinately large number of possibilities. This is exactly the type of problem MCMC was invented to address. Thus, as a practical matter, MCMC together with large computers have made it much easier to be a Bayesian.

"The past 15 years have seen something of a revolution in statistics," wrote Simon Jackman (in 2004), a political scientist who has studied models of the Supreme Court. "The popularization and widespread adoption of MCMC algorithms mean that models and data long relegated to the 'too-hard' basket are now being analyzed."

Data Visualization and Communication

Several speakers at the London workshop touched on issues of the public perception of statistics and the responsibility of professional statisticians to communicate their work effectively. One thought-provoking perspective was given by Mark Hansen, a statistician who is now a professor of journalism at Columbia University. Hansen showed images of some of his art installations that are based on data and statistics. An example is the permanent exhibit "Moveable Type," in the lobby of *The New York Times* tower, in which 560 screens display continually changing snippets of text culled algorithmically from the *Times* article database. This is "communicating statistics" in a form that is more poetic than instrumental: The viewer is provided no explanations, but is presented a view of the newspaper as "data," decomposed and recombined.

David Spiegelhalter spoke about the challenges of explaining risk and uncertainty to the public. Some principles are well established. For example, relative risk (Behavior X will increase your risk of cancer by 50 percent) is perceived differently from absolute risk (Behavior X will change your lifetime risk of cancer from 2 percent to 3 percent). The former sounds more alarming, while the latter sounds like something that people might be willing to live with. Statistical jargon can be a barrier to communication. The public does not understand what a "hazard ratio" is. Spiegelhalter suggested replacing this with a number that people can relate to directly: Behavior X is equivalent to being 8 years older. If the case against smoking had been presented this way, would people perhaps have been quicker to grasp the consequences? Also, visual communication can be very effective. Psychologist Angela Fagerlin found that patients asked to choose between two treatments were susceptible to misleading anecdotes (i.e., Treatment A worked for this patient), even if they had been given statistics that showed the opposite. However, if the statistics were presented visually, the patients retained the information and were effectively "immune" to the misleading anecdote.

Visualization is also central to the work of Hans Rosling (who was not at the London meeting), a Swedish statistician and doctor who has become a YouTube and media star with his multimedia presentations about world demographics. In 2012, he was named as one of *Time's* 100 most influential people. Rosling makes the seemingly dry subject of demographics fascinating with colorful graphics and vivid storytelling. As Rosling said on the BBC program *The Joy of Stats*, "Having the data is not enough. I need to show it in a way that people enjoy and understand." Many other statisticians could learn from his example. For better or for worse, a good visualization is much more convincing to the public than a technically correct report that is full of jargon and numbers. (See, for example, his TED talk at *www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen*.)

1.4 Statistics in Court

Over a 12-year period, from 1984–1995, the Bristol Royal Infirmary in England had an unusually high rate of deaths among infants who underwent open-heart surgery. As early as 1988, an anesthetist complained about operations taking too long, which put the patients at greater risk for death or medical complications. However, it took the death of a baby on the operating table in January 1995 to turn the " Bristol baby case" into a national scandal.

From 1998 to 2001, the British government conducted an official inquiry that eventually cost 14 million pounds and produced a 500-page report. It found systemic failures at Bristol that went beyond the poor performance of one or two surgeons. A number of major changes came as a result of the investigation. Data on the performance of individual surgeons are now publicly available; new standards were set for informing patients about risks and benefits; and Britain formed a permanent healthcare commission, charged with overseeing the quality of care in the National Health Service and at private clinics.

A key ingredient in the report was a statistical estimate of the number of "excess" deaths that had taken place at Bristol. This was difficult to determine, not only because the death rate was subject to large random fluctuations, but also because it was impossible to tell from case records whether a child's death had been caused by surgery or by other factors. On top of that, different patients may have had different degrees of risk. The hospital could have just been unlucky to have a run of sicker babies. Finally, the available data, both at Bristol and at other hospitals, came from various sources and had uneven quality.

In short, the count of excess deaths was fraught with statistical difficulties. Nevertheless, controlling for factors like the patient's age, the type of operation, and the year it was performed, the statisticians estimated that 12 to 34 (out of 41) infant deaths between 1991 and 1995 were excess deaths. We'll never know which babies would have survived at another hospital, but we can confidently say that some of them would have.

David Spiegelhalter, the lead statistician in the Bristol inquiry, was soon called upon in connection with another, even grimmer, case. A general practitioner named Harold Shipman was convicted in 2000 of murdering 15 elderly women by giving them overdoses of opiates. No statistics were involved in this conviction; there was plenty of other evidence, including a fabricated will in which one of the patients left her entire estate to him. However, a subsequent inquiry concluded that Shipman had likely killed at least 215 patients, almost all of them elderly but otherwise in good health, dating all the way back to 1971. This staggering discovery begged the question: Couldn't anything have been done sooner? Couldn't somebody have seen that the mortality rate of this doctor's patients was unacceptably high?

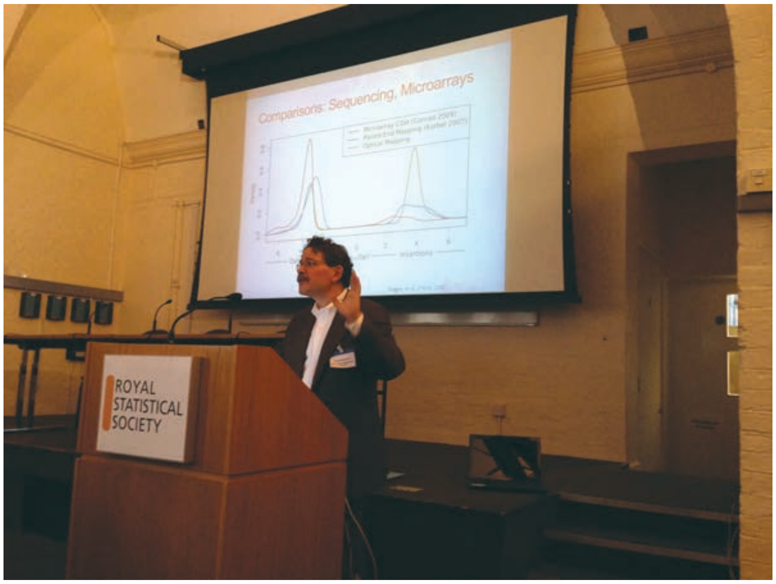
The same question had arisen after the Bristol case. To answer it, Spiegelhalter adapted sequential analysis techniques that were developed in World War II to monitor industrial processes, such as munitions production. He concluded that, had the data been available and had the right statistical methods been in place, Shipman's practice could have been identified as suspicious long before 1998, when he was finally caught.

However, it is important to use an abundance of caution when extrapolating from statistics to malfeasance. This was the salient lesson of another case that received widespread media attention, the case of the Dutch nurse Lucia de Berk.

The bizarre case of de Berk started in similar fashion to the Bristol baby scandal—with the unexpected death of one infant, named Amber, on September 4, 2001, at the Juliana Children's Hospital in The Hague. De Berk was a nurse on duty at the time. Some of her co-workers had noticed previously that she had been on duty during a suspiciously large number of unanticipated patient deaths, or "incidents." The hospital director investigated the records of two previous wards she had worked in and noticed that there had been several incidents there as well. The hospital reported the evidence to the police, alleging that de Berk had committed five murders and five attempted murders.

The human brain is unfortunately quite prone to seeing patterns where there is only randomness. For example, if you shake up a jar of jelly beans of different colors, you will likely see surprisingly large clumps of beans of the same color. Before concluding that the beans of the same color are stuck together, or before concluding that the cluster of deaths on de Berk's watch were related, one should first do a careful statistical analysis to see whether such a cluster could be explained by chance.

Instead, the police consulted a lawyer who was not a professional statistician and who performed a very shoddy analysis. He claimed that there was only 1 chance in 342 million that so many deaths would have happened during de Berk's shifts by chance alone. Though he later backed away from the claim, the number 342 million took on a life of its own. De Berk was convicted of four counts of murder in 2003 and sentenced to life in prison. The conviction was upheld by the Dutch Supreme Court in 2006, although the court took pains to state that the verdict was based on physical evidence and *not* based on statistical reasoning—most likely because serious doubts had already arisen about the latter.



David Schwarz looks at applications of statistics in genetics.

Statisticians Richard Gill and Willem van Zwet pointed out numerous statistical flaws behind the notorious "1 in 342 million," which amounted (in Gill's words) to an "abuse of every basic rule in the statistics textbook." One flaw was technical, but bears mentioning because it is such a common misconception. The "p-values" (see §1.1) that are obtained in conventional significance tests are *not* probabilities in the usual sense. Yet the consultant had combined them as if they were. To illustrate, suppose there were a 10 percent probability of x or more deaths occurring in Ward A and a 10 percent chance of y or more deaths occurring in Ward B during de Berk's shifts. Then one would think there was only a 1 percent (i.e., 10 percent times 10 percent) chance that both events would happen. But if the "10 percents" are p-values, not probabilities, then they are just numbers attached to data. A more relevant question would be, "What are the chances of picking two random numbers between 0 and 1 that multiply to less than 1/100?" *This* probability is more than 5 percent—considerably more than the naïve estimate. In de Berk's case, Gill showed that the correct probability should have been calculated as 1 in 100,000—not 1 in 342 million.

That still sounds pretty damning, but there were many other problems with the data that actually make the calculation moot. The deaths in the other wards would not have been noticed if the hospital administration hadn't gone looking for them; this is called "confirmation bias." In addition, data on all the times a death had *not* happened on de Berk's shifts was overlooked ("selection bias"). In view of the biased way the data were acquired, a professional statistician would have to say that no conclusions could be drawn from them. The calculations discussed above would be valid only if the data came from a monitoring system set up *in advance* that treated every nurse equally.

In 2010, de Berk's case was reopened and her conviction was overturned. Once again, the court said that statistics were not the issue. There was new medical evidence that the poison de Berk allegedly used (digoxin) can accumulate in the body naturally after death, so its presence in the alleged victims' bodies did not indicate foul play. However, the flimsiness of the statistical case against her may well be the reason that doctors took a closer look at the medical evidence and found that wanting as well.

A strong message emerges from these three cases, even though the details are different. Statistical reasoning *can* be useful for spotting nefarious behavior. But they can do so only if the data are gathered in a planned, methodical, and unbiased way. In addition, the data must be analyzed by competent professionals who understand the assumptions behind the statistical models. The best defense against "lying with statistics" (even inadvertently) is … to hire a statistician.

1.5 Statistics, Genomics, and Cancer

In the 1970s, when U.S. President Richard Nixon declared a "war on cancer," the disease was seen as a monolithic adversary; treatment was harsh and successes were few. But during the 1980s and 1990s, cancer researchers discovered that cancers are almost as unique as the patient. There isn't just one disease called "breast cancer"—there are many. Each kind of breast cancer has a different prognosis and calls for different kinds of treatment. It took 20 years for scientists to recognize and find a treatment for estrogen-receptor positive cancer, from the discovery of the receptor gene (called HER2) in 1978 to the FDA's approval in 1998 of Herceptin.

But a new wind was blowing in cancer research and in all of medicine by the end of the 1990s. It was the era of genomics. The invention of the microarray, or "gene chip," made it possible for scientists to study gene expression in cells (including cancer cells) not just one gene at a time, but thousands at a time.

One kind of gene chip contains short snippets of DNA from known genes, which are attached to a glass base. Often, they are arranged in a rectangular array, like pixels in a photograph. The rows and columns in the array have meanings. For instance, the rows may represent samples from different cancers and the columns might represent genes. When a

DNA sample from a patient is inserted into the chip, it will attach to those snippets that match some subsequence of DNA in the sample. When this happens, fluorescent chemicals in the microarray cause that dot to light up. The array displays a pattern of bright red and green lights, like a Christmas tree, that identify which genes are present in that sample of DNA.

With microarrays, biologists could test for hundreds or thousands of genes simultaneously. Could the microarray data be used to identify other previously unsuspected kinds of cancer, analogous to the estrogen receptor-positive variants?

The answer was yes, thanks to a statistical technique called clustering and a visualization technique called heat maps. The idea is to look for rows or columns in the microarray with similar patterns of red and green lights. The most similar rows are placed in a cluster. Then the process is repeated. The clusters that are most similar are grouped into even larger clusters. The process can be continued until everything is placed into one big cluster, or it can be stopped at an earlier stage when there are just a few big clusters.

After suitable clusters have been found, the rows and columns of the microarray data can be rearranged so that adjacent rows or columns are in the same cluster. As simple as it sounds, this makes an amazing difference. An initially jumbled pattern of reds and greens will take on a remarkable plaid appearance after it is converted to a heat map. The clusters—the plaid bands—literally pop out of the picture; you don't have to be an expert to see them. However, two things should be pointed out about heat maps. First, they are only a visualization technique and only as valid as the clustering algorithm that produces them. Second, in spite of the belief of genomics researchers that heat maps are a recent invention, statisticians have been producing images like this for almost 100 years!

In a landmark paper in 2000, a large team of researchers led by geneticist David Botstein applied cluster analysis to malignant breast tumors and found that they could be classified into five distinct groups. What was interesting about this was that only four of the groups were already known. One of them, for example, was the HER2-positive group. Botstein called the newly discovered group of breast cancers "basal-like," because the gene expression pattern was similar to cells in the basal (outer) layer of the breast. It was the first time that a statistical program had discovered a "biomarker" for a distinct subtype of cancer.

Of course, that was only the beginning of the story. One of the drawbacks of cluster analysis is that it will produce clusters whether actual meaningful groups exist or not. The findings had to be replicated and it had to be shown that the new cluster was biologically relevant, not just an artifact of the clustering algorithm.

Official/Government Statistics

It is a little-known fact that the word "statistics" actually comes from the root "state"—it is the science of the state. Thus, government or official statistics have been involved in the discipline from the beginning, and, for many citizens, they are still the most frequently encountered form of statistics in daily life.

Several trends are placing new demands on official statisticians. Many governments are moving toward open government, in which all official data will be available online. Many constituents expect these data to be free. However, open access to data poses new problems of privacy, especially as it becomes possible to parse population data into finer and finer units. Free access is also a problem in an era of flat or declining budgets. Though information may want to be free, it is certainly not free to collect and curate.

At the same time, new technologies create new opportunities. There are new methods of collecting data, which may be much cheaper and easier than traditional surveys. As governments move online, administrative records become a useful and searchable source of information. Official statisticians will face a Big Data problem similar to private business as they try to figure out what kinds of usable information might exist in these large volumes of automatically collected data and how to combine them with more traditionally collected data. They also need to think about the format of the data; mounds of page scans or data that are presented out of context may not be very useful. With proper attention to these issues, both old democracies and new democracies can become more transparent, and the citizens can become better informed about what their governments are doing.

In the case of basal-like breast cancer, the results in the ensuing years have been clear. It is a distinct entity from other breast cancers. It is particularly prevalent in younger women and in black women and is associated with a gene (BRCA1) that can be identified by a commercially available test. Clinically, it has one of the poorest prognoses of any breast cancer, because it is aggressive and because it is typically "triple-negative"—it does not display any of the three common estrogen receptors. That means that drugs like Herceptin, which target an estrogen receptor, are ineffective.

In short, basal-like breast cancer is exactly the kind of cancer that we most need a treatment for, a particularly vicious cancer that especially hits younger women. But it's hard to base a treatment on a negative test. The value of the microarray study is that it identifies a positive criterion for identifying these cancers and suggests genetic targets that a drug might be able to exploit. As of this writing, drugs for basal-like breast cancer have been fast-tracked by the FDA. Though it has been 14 years since the discovery of basal-like breast cancer, such a development time is normal, or even quick, in the context of cancer research.

The discovery of basal-like breast cancer is a model of what biologists hope to accomplish through genome research, as well as the ways in which statisticians can contribute. However, in all fairness, it must be pointed out that this is not yet a typical example. The research literature is full of putative biomarker discoveries. Yet, according to a 2012 article by oncologist Scott Kern, "Less than 1% of published cancer biomarkers actually enter clinical practice." (He defined "entering clinical practice" to mean that a patient can actually get a test for that biomarker that will be reimbursed by an insurance company.) In some cases, a biomarker may be valid but not be useful for clinical treatment. In other cases, the supposed discoveries are simply wrong and cannot be replicated. The difficulty of reproducing published results has been much discussed in recent years and will be discussed later in this report.

1.6 After the Gold Rush: Kriging and Geostatistics

Not too many people get to see their names turned into a verb. But in the early 1950s, a South-African mining engineer named Danie Krige revolutionized the mining industry to such an extent that his name has been affixed to a statistical technique that he helped to invent. "Kriging" now refers to a method for interpolating data collected at sparse sample points in a way that minimizes the expected error of the estimates.

The application Krige developed his method for was gold mining. The huge gold deposits in the Witwatersrand (the world's richest gold field) are buried deep underground, so miners have to drill boreholes to figure out where the highest-grade ore is. But the boreholes are just isolated points in a landscape of thousands of square kilometers. Until Krige, there had been no formal statistical method for estimating the grade of ore between boreholes.

Krige made three simple assumptions. (Though they are debatable, some assumptions *always* have to be made to derive any kind of mathematical or statistical model. Often, alternative models are applied to determine the robustness of the conclusions to specific assumptions.) First, he assumed that there was some average concentration of gold throughout the gold fields. Second, there are random deviations from this average, due to all the churning and scrambling of rock through Earth's geologic history. Third and most important, the deviations are *correlated*. A borehole drilled in one spot can tell you about the ore concentration nearby. The form of the correlation is not specified in advance, but it is assumed to remain unchanged or vary slowly throughout the field. For example, if the strata that the gold lie in are oriented north-to-south, then the correlations may be greater in that direction than in the east-west direction.

The method Krige derived from these assumptions was quintessentially statistical, because the prospector pays more attention to the covariances (a measure proportional to the correlation of the random deviations) than the actual values of the borehole measurements at first. This information is summed up in a graph called a "variogram." Once the variogram is known, an estimate of the grade of ore at any point in the field can be interpolated by taking a weighted average of the grade at the nearest boreholes. The weights are computed from the variogram by a formula that Krige developed. The resulting estimate is the *best linear unbiased predictor* of the actual ore concentration.

Interpolation techniques of various kinds were developed long before Krige. However, none of these approaches is inherently statistical. They are not based on a model of uncertainty. Instead of minimizing the expected error in the predictor, they optimize other things—the smoothness of the interpolated function, for instance. But a miner doesn't care whether the predicted distribution of ore is smooth. He just wants to make sure it's correct!

Ironically, Krige himself did not fully understand the optimality of his technique. It was Georges Matheron, a French mathematician and geologist, who clarified the theory in the 1960s and introduced the name "kriging." Matheron also developed a number of alternative methods for cases where Krige's simple model is not reasonable. For instance, in "ordinary" kriging (invented by Matheron), you don't assume that you know the average concentration of ore in the field. In "universal" kriging (also invented by Matheron), you can assume that the average ore concentration has a distinct trend. For example, it may gradually increase in the north-south direction.

New variants are being discovered even today. Matheron's techniques—like many traditional statistical techniques—don't scale well. In practice, if you have more than a few hundred data points, your calculation will grind to a halt (or a slow crawl). To solve this problem, "fixed rank kriging," introduced in 2006, brings kriging into the era of Big Data. Krige's and fixed-rank kriging speeds up the algorithm by orders of magnitude by assuming a particularly parsimonious form for the variograms.

The utility of kriging goes far beyond mining, although it is not always called by that name in more distantly related fields. Krige's statistical model was discovered in the Soviet Union before it was known in the West, and a Soviet meteorologist named Lev Gandin independently proved Matheron's theorems about the optimality of kriging. The Russians called the variogram a "homogeneous structure function" and kriging "optimal interpolation," or OI. Unfortunately, they weren't as clever at naming things as Georges Matheron was!

Questions of nomenclature aside, the applicability of kriging to meteorology is not hard to explain. Meteorologists constantly have to produce maps with smooth curves from data that are concentrated at a small number of observation points. While kriging may be too computationally intensive to use for a daily weather map, it is perfect for things like estimating the thickness of the snowpack or drawing a map of the ozone hole.

In recent years, many farmers have adopted "precision agriculture," a development that owes as much to kriging as to new technology. Farmers can now manage their crops in a way that takes account of variations in growing conditions *within* a field. They measure statistics such as soil acidity or productivity of past crops at a few locations and use kriging to create a smooth map of the whole field. The map lets them decide where to put lime or fertilizer and how much. By limiting the use of these chemicals, they not only save money, but also minimize damage to the environment.

Though developed for geosciences, kriging can even be extended to outer space. One way that astronomers infer the distribution of matter in the universe involves what they call the "Lyman-alpha forest." Light from extremely distant quasars passes through many galaxies and clouds of gas on its way to our telescopes here on Earth. A specific wavelength of this light (the Lyman-alpha band) is absorbed by the hydrogen atoms it encounters on the way. The amount of absorption gives astronomers a clue to how much matter is out there. But the information is limited to one dimension—the straight line of sight between Earth and the quasar. That line of sight is like a borehole in a gold field, only it's a borehole billions of light-years long. Unfortunately, there are not many quasars, and there is a lot of empty space between the boreholes (or the trees in the Lyman-alpha forest, to use the metaphor that the astronomers prefer). Not surprisingly, kriging is one method to infer the distribution of matter in those in-between regions.

1.7 'Analytics' in Sports and Politics

In recent years, statistics and statistical methods have achieved unprecedented prominence in the sports world, where they are often referred to as "analytics." The best-publicized example was the adoption of analytics by the Oakland (California) Athletics baseball team in the late 1990s and early 2000s, the subject of the book (and movie) *Moneyball*. The Athletics enjoyed—and are still enjoying—a run of success that is incommensurate with their financial resources. *Moneyball* attributes their success to the visionary general manager, Billy Beane, who was willing to adopt unconventional metrics of baseball ability. (A general manager makes hiring and trading decisions and is thus responsible for the personnel on a baseball team.)

In recent years, statistics and statistical methods have achieved unprecedented prominence in the sports world, where they are often referred to as "analytics."

The analytics movement also received a boost from the success of the Boston Red Sox. Though not a small-market team in the mold of the Athletics, the Red Sox franchise had long been considered to be "cursed" (not a statistical or scientific concept!) because of its inability to win a World Series since 1918. In 2002, the Red Sox hired a new analytics-oriented general manager, Theo Epstein, and they brought in the founder of baseball analytics (also known as "sabermetrics"), Bill James, as a consultant in 2003. Sure enough, in 2004, they won their first World Series in 86 years and followed it with two more titles in 2007 and 2013. Analytics suddenly seemed like a golden path to success. Today, according to Oakland Athletics director of baseball operations Farhan Zaidi, every professional baseball team has an analytics department of some kind.

The sabermetric movement really began in the early 1980s, when James, in his annual *Baseball Abstract*, began subjecting baseball statistics to quantitative scrutiny. Baseball had long been one of the most data-rich sports, but James found that many of the traditional measures of success, such as "batting average" for hitters or "won-loss record" for pitchers, had little predictive value. Batting average arbitrarily ignored an important way a hitter can contribute to his team (by drawing a walk, thus reaching base without actually hitting the ball). Batting average is also biased in favor of hitters in "hitter-friendly" parks. For pitchers, wins and losses seem like a natural metric, but they are contaminated by many

Statisticians Were Here: A Word on Terminology

As used in this report, the words "data" and "statistics" mean two different things. This point, which is so evident to statisticians that it barely needs mentioning, is nevertheless crucial to establish at the outset. For the general public, the two words are nearly synonymous, and most people would have great difficulty explaining the difference if pressed to do so. The confusion is very important, because it leads to a lack of understanding of or appreciation for what statisticians do.

In this report, the word "data" means a series of measurements or observations—usually in numerical form, but not necessarily—of some phenomenon. On the other hand, the word "statistics" refers to an academic discipline and a set of best practices to convert data into meaningful, actionable information about the real world, *particularly in the presence of uncertainty*. The word "statistic" is also used in the specialist literature to mean "a numerical summary of data." For example, the median and the mean are both examples of a statistic.

For example, a statistician may be given data on the number of unique visitors to a website each day. The goal is to transform the data into insights, such as this:

The number of visitors tends to be greater on days when there is a new post, or the number of visitors behaves like a linear function of the advertising budget, plus a certain amount of random variation.

Do the data support such a conclusion? If so, how strongly? To evaluate these statements, the statistician may compute certain kinds of statistics, such as means and linear regression coefficients and *p*-values. (The latter term will be mentioned in several places in this report. In brief, *p*-values are a way of assessing the "statistical significance" of a difference between two groups, such as "days with a post" and "days without a post.") No matter what conclusion the statistician reaches, there will always be uncertainty about the result. Statistics is meant to quantify uncertainty, not to hide it.

Uncertainty comes in two flavors: random and systematic. Both are of concern to statisticians. They have developed a suite of powerful mathematical tools over the years to estimate the size and nature of random uncertainty in various contexts. This is the professional expertise that statistics students acquire during their education and that scientists call upon every day. Systematic uncertainty is just as important, and not as easy to manage. Understanding systematic uncertainty requires a certain skeptical frame of mind, which is prepared to look for hidden biases in data collection and to reject the data outright if the biases cannot be corrected for.

This clear-eyed skepticism can be taught in school, but it is also learned through experience and example. It is one of the most important ingredients that professional statisticians can bring to the table in scientific research, ideally even before the data are collected. "The statistician who supposes that his main contribution to the planning of an experiment will involve statistical theory, finds repeatedly that he makes his most valuable contribution simply by persuading the investigator to explain why he wishes to do the experiment," wrote Gertrude M. Cox, a pioneering American statistician and former president of the American Statistical Association.

In short, statistics is a profession that is built on data, but statistics is more than data. The data should be collected with a purpose (although appropriate secondary use of data, different from the originally intended use, can also be informative). Attention should be paid to identifying confounding factors and removing systematic sources of bias. When conclusions are drawn, they should be presented in a way that acknowledges the uncertainty and estimates its size. When all of these ingredients are present, whether the context is economics or biology or Web commerce, it is like a signature: "Statisticians were here."

confounding factors. A pitcher can pitch poorly, but nevertheless get credit for a win because his team scores many runs. Likewise, a powerful pitching effort can go to waste for reasons beyond the pitcher's control. At best, wins and losses are an imperfect measure of a pitcher's value, and they are not causative. A good won-lost record is a *result* of a pitcher's success; it is not a predictor.

These principles—controlling for confounding factors, eliminating bias, distinguishing correlation from causation—have been part of good statistical practice for a long time. However, despite the abundance of *data* in baseball, the data had never been subjected before to rigorous *statistical thinking*. It is not surprising that early adopters of sabermetrics enjoyed considerable success against teams that used metrics with no sound statistical foundations.

In 2007, Nate Silver, a well-known sabermetrician, ventured into another high-profile area: politics. Frustrated by the lack of solid methodology among many political pundits, he began writing a blog called FiveThirtyEight.com. In the 2008 U.S. presidential election, he correctly forecasted the results of 49 out of 50 states, as well as all 35 Senate races. This seemingly uncanny success rate attracted a great deal of media attention, and his blog was acquired by *The New York Times*. In the 2012 presidential election, he correctly called the results of all 50 states and 31 out of 33 Senate races. A week before the election, he gave President Obama a 70 percent chance of winning, and by the day of the election, the probability was up to 90 percent, even while many commentators were calling the election a toss-up.

Silver does not rely (at least not heavily) on inside information. His method is simply to aggregate existing polls, weighting them in the most informative way. Even a poll that is habitually biased toward one party may contain useful information. It may have access, for whatever reason, to potential voters that other polls miss. An aggregate of polls can incorporate more information than any individual poll, and in the end, it will nearly always outperform individual polls.

To some extent, the remarkable feat of correctly predicting 50 out of 50 states obscures what Silver really did and why his methods work. As he explains in his bestselling book *The Signal and the Noise*, there is a difference between predicting and forecasting. A prediction is a single outcome: "Obama will win Ohio." A forecast is a probability statement: "Obama has an 80 percent chance of winning Ohio."

An example of forecasts that we are all familiar with is a weather forecast. As Silver points out, we are better at forecasting the weather than almost any other uncertain phenomenon in our lives: stock markets, earthquakes, terrorist attacks. There are many reasons, but one of them is that weather forecasts *do not pretend* to be predictions. They always come with a probability of error. If an honest weather forecaster says there is an 80 percent chance of rain, he should be wrong one-fifth of the time. It should literally rain 80 percent of the times he issues such a forecast and it should not rain the other 20 percent. The data show, in fact, that this is exactly what happens with weather forecasts.

U.S. presidential elections are especially suitable for a probabilistic approach because they involve 51 mini-elections (including the District of Columbia) that are correlated in a complicated way. An expert who relies solely on experience and intuition cannot assess the probabilities well enough, but a computerized model can.

It is exceedingly important for a political forecaster to think like a weather forecaster in terms of a probability distribution with a certain range of error. Of course, this goes against the inclination of most people reading polls. Human nature does not like uncertainty. We like predictions: "President Obama will win the election." But a forecast is more honest, and more trustworthy in the long run, if it makes some explicit statement about the range of possible outcomes.

Like the successes of Theo Epstein and Billy Beane, Nate Silver's success was not achieved in a vacuum. Statisticians have been forecasting things like voter turnout and the effects of gerrymandering for at least three decades, and Silver's methods are not particularly novel or better than the others. What is new is the amount of media attention that he has attracted. He has gone beyond the traditional academic journals, disseminating his research via the much more widely read medium of the Internet. In this way, he has greatly enhanced the public profile of statistics. This approach clearly has its dangers, and it cannot substitute for peer-reviewed publication in journals. However, it does provide an opportunity and a model for public engagement that other statisticians might think about emulating. Not only that, it shows that sound, principled statistical reasoning does have a chance to be heard amidst the Babel of conflicting opinions that is the Internet. ❖

SECTION 2.
Current Trends and Future Challenges in Statistics: Big Data

Without a doubt, the most-discussed current trend in statistics at the Future of Statistics Workshop was Big Data. The ubiquity of this phrase perhaps conceals the fact that different people think of different things when they hear it. For the average citizen, Big Data brings up questions of privacy and confidentiality: What information of mine is out there, and how do I keep people from accessing it? For computer scientists, Big Data poses problems of data storage and management, communication, and computation. And for statisticians, Big Data introduces a whole different set of issues: How can we get usable information out of databases that are so huge and complex that many of our traditional methods can't handle them?

At the workshop, all perspectives were present, from "Big Data is an opportunity that statisticians should not miss" to "Big Data will change statistics as we know it" to "Big Data is just a fad and we should not buy into the hype." This section of the report will not endorse any of these points of view, but will summarize the state of play and the challenges that Big Data poses to statistical science.

2.1 Examples of Big Data

Some of the most commonly cited forms of Big Data are:

- Commercial databases, such as those of Google or Facebook.
- Government or official data.
- Human genome databases. Even a single human genome contains more than 3 billion base pairs. The 1000 Genomes Project, for example, has now collected 200 terabytes (200 trillion bytes) of data.
- Human brain data. A single human brain scan consists of data on more than 200,000 voxel locations, which could be measured repeatedly over 300 time points. The Human Connectome Project, which is gathering MRI images of 1,200 patients over a five-year period, has publicly released 7 terabytes of data as of early 2014.
- Databases in physics and astronomy. For example, the Large Hadron Collider experiment generates more than a petabyte (1,000 trillion bytes) of data per year. The Large Synoptic Survey Telescope (LSST), which is scheduled to become operational in 2020, will generate a petabyte of data per night.

2.2 Not Just Big, but Different

For statisticians, Big Data challenges some basic paradigms. One example is the "large *p*, small *n*" problem. (Comment on notation: This "*p*" is the number of variables, not a *p*-value!) Classical statistics provides methods to analyze data when the number of variables *p* is small and the number of data points *n* is large. For example, a biologist might want to estimate how the population of a particular species of fish in a river varies, depending on variables like the depth of the river, the size of the watershed, the oxygen content of the water, and the water temperature. These are four variables, and the data might be taken from, say, 50 or 100 streams. The sample size *n* far exceeds the number of variables *p*.

In some (though not all) Big Data applications, this situation is reversed. In genomics, the researcher might collect data on 100 patients with cancer to determine which genes confer a risk for that cancer. Unfortunately, there are 20,000 genes in the human genome and even more gene variants. Genome-wide association studies typically look at a half million "SNPs," or locations on the genome where variation can occur. The number of variables (p = 500,000) is much greater than the sample size (n = 100). Similarly, in neuroimaging studies, the variables correspond to voxels or regions of interest, which often outnumber the participants in a survey.

In either situation, the goal is to develop a model that describes how an outcome variable (e.g., size of the population of fish, presence of cancer) is related to the p other variables (and to determine which variables are important in characterizing the relationship). The model depends on parameters, one for each variable, that quantify the relationship, and fitting the model to data involves estimating the parameters from data and assessing the evidence that they are different from zero (reflecting an important variable). When p is larger than n, the number of parameters is huge relative to the information about them in the data. Thousands of irrelevant parameters will appear to be statistically significant if one uses small-data statistics. In classical statistics, if the data contain something that has a one-in-a-million chance of occurring, you can be confident that it isn't there by chance. But if you look in a half million places, as in the Big Data world, suddenly it isn't so unusual to make a one-in-a-million discovery. Chance can no longer be dismissed as an explanation. Statisticians call this the "look-everywhere" effect, and it is a major problem with data-driven, rather than hypothesis-driven, approaches to science.

Statisticians have already found several good ways to deal with the look-everywhere effect. Most data sets have only a few strong relationships between variables, and everything else is noise. Thus most of the parameters simply should not matter. One way to make them not matter is to assume all but a few parameters are equal to zero. Some technical advances in the last few years have made this a particularly promising way to extract a needle of meaningful information from a haystack of data. One such technique is called L_1-minimization, or the LASSO, invented by Robert Tibshirani in 1996. Coincidentally, L_1-minimization was discovered almost simultaneously in the field of image processing, where it enables the extraction of an image in sharp focus from blurry or noisy data.

Another widely applied technique, especially in genome and neuroimaging research, is the false discovery rate (FDR) proposed by Yoav Benjamini and Yosi Hochberg in 1995. This is a method of providing an alternative interpretation to statistical significance tests to take account of the look-everywhere effect. For example, if a study finds 20 SNPs with a statistically significant association with cancer, and it has a false discovery rate of 10 percent, then you should expect two of the 20 "discoveries" to be false, on average. The FDR doesn't tell you which ones (if any) are spurious, but that can sometimes be determined by a follow-up study.

These examples make it clear that Big Data should not be viewed only as a challenge or a nuisance. It is also an opportunity for statisticians to re-evaluate their assumptions and bring new ideas to the forefront.

Financial Statistics

The financial crisis of 2007 and 2008, and the recession that followed, brought to light a number of fundamental flaws in the evaluation of risk. The Basel accords required financial institutions to report the risk of their investments. However, the metric that became standard in the community (Value at Risk) was merely a measurement of convenience that had been developed by one company and had little statistical theory behind it. Also, the methods that were used to value credit derivatives (such as the Gaussian copula) were deeply flawed and likewise became popular simply because "everybody else is doing it." In particular, these methods assumed that certain events, such as the default of a homeowner in Las Vegas and the default of a homeowner in Miami, are independent, or at most, weakly dependent. In normal times, this is a harmless assumption. However, during a financial panic, such events become highly correlated. Some statisticians sounded a warning long before 2007 about the immensely risky positions that were being passed off as sound, but they were ignored.

It remains to be seen whether financial institutions can learn to police themselves, but it's fair to say that if a solution exists, it will require an approach that owes less to herd psychology and more to sound statistical principles.

2.3 Big *n*, Big *p*, Little *t* (time)

As both n and p grow, statisticians have to grapple with another difficulty that they have never had to take seriously before: the pressure of time.

Classical statistics was always done in an offline mode; a lot of the theory was developed in an era (the early 1900s) when "online" didn't even exist. The researcher collected the data, then went back to his or her office and analyzed it. The time for analysis was essentially unlimited. The data were small anyway, so the statistician never had to think about whether the calculations were done in an efficient way.

However, in the era of Big Data, things are different. Web companies make their money by trying to predict user reactions and elicit certain user behaviors (such as clicking on an advertisement sponsored by a client). To predict the response rate, they need a statistical model with large n (millions of clicks) and large p (tens of thousands of variables—which ad to run, where to put it on the page, etc.). In this case, n is much larger than p, so classical statistical techniques could apply in theory.

However, the algorithms needed to perform a regression analysis don't scale well. The Web company might have only milliseconds to decide how to respond to a given user's click. Not only that, the model constantly has to change to adapt to new users and new products. The Internet is like a massive ongoing experiment that is constantly changing and never finished.

To address the problem of time, statisticians have started to adopt and adapt ideas from computer scientists, for whom speed has always been an issue. The objective in some cases may not be to deliver a perfect answer, but to deliver a good answer fast.

Yet at the same time, statisticians cannot afford to stop thinking like statisticians. They bring a type of expertise to the corporation that computer scientists (for the most part) can't: Statisticians understand uncertainty. Where a computer scientist sees a number, a statistician sees a range of possibilities. Predictions get better when they are thought of as forecasts, which have an inherent uncertainty. Statisticians are also uniquely qualified to make inferences—to abstract connections between variables from data and determine which connections are real and which are spurious.

One way to make statistical procedures more efficient is to parallelize them—to write algorithms that can run on many computers or many processors at once. At the London workshop, Michael Jordan presented a case study of one such application called the Bag of Little Bootstraps (BLB). The "bootstrap" is a standard method, invented by Bradley Efron in 1979, for inferring the probability distribution of a population from a sample. Like so many statistical methods, it is computationally intensive. However, it is an ideal candidate for parallelization, because it involves generating numerous independent rounds of simulated data. In 2012, Jordan and a group of colleagues ran the BLB on the Amazon cloud computing platform using various data sets in the public domain. The BLB generated results comparable to the regular bootstrap, but orders of magnitude faster. The supporters of the research read like a Who's Who of high-tech companies: Amazon, Google, SAP, Cisco, Oracle, and many more. It's fair to say that Silicon Valley has noticed the need for new statistical tools to deal with Big Data.

2.4 New Data Types

Another current trend that has emerged in tandem with Big Data is new data types. These data are not simple numbers; they come in the form of a function, image, shape, or network. For instance, first-generation "functional data" may be a time series, measurements of the blood oxygenation taken at a particular point and at different moments in time. In contrast to traditional scalar or multivariate data, the observed function in this case is a sample from an infinite-dimensional space (because it involves knowing the oxidation at infinitely many instants). Such "infinite-dimensional" data already demand more sophisticated methods.

But even that isn't the end of the story. At the London workshop, Jane-Ling Wang talked about next-generation functional data, which include correlated random functions. The functions may have conventional numerical values, or the values may be images or shapes. For instance, the observed data at time *t* might be the region of the brain that is active at time *t*.

Brain and neuroimaging data are typical examples of next-generation functional data, and they are the focus of two recent research initiatives on human brain mapping, one by the Obama administration and another by the European Union. These projects aim at mapping the neuron activity of the whole human brain and understanding how the human brain works.

Next-generation functional data are invariably Big Data, but they are not just big, they are also complex. They require the invention or importing of ideas from areas of mathematics outside what is conventionally thought of as statistics, such as geometry (to describe abstract shapes) or topology (to describe the spaces that the data are sampled from). These are examples of the ways in which Big Data not only challenge, but also enrich, the practice of statistics.

2.5 Privacy and Confidentiality

The year 2013 was the year when many Americans woke up to the volumes of data that are being gathered about them, thanks to the highly publicized revelation of the National Security Agency's data-mining program called PRISM. In this climate, public concerns about privacy and confidentiality of individual data have become more acute. It would be easy for statisticians to say, "Not our problem," but, in fact, they can be part of the solution.

Two talks at the London workshop, given by Stephen Fienberg and Cynthia Dwork, focused on privacy and confidentiality issues. Fienberg surveyed the history of confidentiality and pointed out a simple, but not obvious, fact: As far as government records are concerned, the past was much worse than the present. U.S. Census Bureau records had no guarantee of confidentiality at all until 1910. Legal guarantees were gradually introduced over the next two decades, first to protect businesses and then individuals. However, the Second War Powers Act of 1942 rescinded those guarantees. Block-by-block data were used to identify areas in which Japanese-Americans were living, and individual census records were provided to legal authorities such as the Secret Service and Federal Bureau of Investigation on more than one occasion. The act was repealed in 1947, but the damage to public trust could not be repaired so easily.

And for statisticians, Big Data introduces a whole different set of issues: How can we get usable information out of databases that are so huge and complex that many of our traditional methods can't handle them?

There are many ways to anonymize records after they are collected without jeopardizing the population-level information that the census is designed for. These methods include adding random noise (Person A reports earning $50,000 per year and the computer adds a random number to it, say –$10,000, drawn from a distribution of random values); swapping data (Person A's number of dependents is swapped with Person B's); or matrix masking (an entire array of data, p variables about n people, is transformed by a known mathematical operation—in essence, "smearing" everybody's data around at once). Statisticians, including many at the U.S. Census Bureau, have been instrumental in working out the mechanics and properties of these methods, which make individual-level information very difficult to retrieve.

Cryptography is another discipline that applies mathematical transformations to data that are either irreversible, reversible only with a password, or reversible only at such great cost that an adversary could not afford to pay it. Cryptography has been through its own sea change since the 1970s. Once it was a science of concealment, which could be afforded by only a few—governments, spies, armies. Now it has more to do with protection, and it is available to everyone. Anybody who uses a bank card at an ATM machine is using modern cryptography.

One of the most exciting trends in Big Data is the growth of collaboration between the statistics and cryptography communities over the last decade. Dwork, a cryptographer, spoke at the workshop about differential privacy, a new approach that offers strong probabilistic privacy assurances while at the same time acknowledging that perfect security is impossible. Differential privacy provides a way to measure security so that it becomes a commodity: A user can purchase just as much security for her data as she needs.

Statisticians not only know how to ask the right questions, but, depending on the answers, they may have practical solutions already available.

Still, there are many privacy challenges ahead, and the problems have by no means been solved. Most methods of anonymizing do not scale well as *p* or *n* get large. Either they add so much noise that new analyses become nearly impossible or they weaken the privacy guarantee. Network-like data pose a special challenge for privacy because so much of the information has to do with relationships between individuals. In summary, there appears to be "no free lunch" in the tradeoff between privacy and information.

2.6 Quality of Data

One of the underrated services that statisticians can provide in the world of Big Data is to look at the quality of data with a skeptical eye. This tradition is deeply ingrained in the statistical community, beginning with the first controlled trials in the 1940s. Data come with a provenance. If they come from a double-blind randomized controlled trial, with potential confounding factors identified and controlled for, then the data can be used for statistical inference. If they come from a poorly designed experiment—or, even worse, if they come flooding into a corporate web server with no thought at all given to experimental design—the identical data can be worthless.

In the world of Big Data, someone has to ask questions like the following:

- Are the data collected in a way that introduces bias? Most data collected on the Internet, in fact, come with a sampling bias. The people who fill out a survey are not necessarily representative of the population as a whole. (See sidebar, "Why Bias Matters.")
- Are there missing or incomplete data? In Web applications, there is usually a vast amount of unknown data. For example, the movie website Netflix wanted to recommend new movies to its users using a statistical model, but it only had information on the handful of movies the user had rated. It spent $1 million on a prize competition to identify a better way of filling in the blanks.
- Are there different kinds of data? If the data come from different sources, some data might be more reliable than others. If all the numbers get put into the same analytical meat grinder, the value of the high-quality data will be reduced by the lower-quality data. On the other hand, even low-quality, biased data *might* contain some useful information. Also, data come in different formats—numbers, text, networks of "likes" or hyperlinks. It may not be obvious to the data collector how to take advantage of these less-traditional kinds of information.

Statisticians not only know how to ask the right questions, but, depending on the answers, they may have practical solutions already available.

2.7 A Statistician by Any Other Name …

"A data scientist is someone who can compute better than any statistician, and who can analyze data better than any computer scientist." – Josh Wills (director of data science, Cloudera)

"A data scientist is a statistician who also understands those parts of computer science and mathematics concerned with data manipulation." – David Hand (statistician, Imperial College London)

"A data scientist is a statistician who is useful." – Hadley Wickham (statistician, Rice University)

For all the reasons listed in the last section, statisticians can offer tremendous value to organizations that collect Big Data. However, statisticians at the Future of Statistics Workshop were concerned that they, or more precisely their students, might get shut out. The job openings in Silicon Valley that their students are applying for are not designated for "statisticians"—they are for "data scientists." The graduate students want these jobs. Anecdotally, students see a job at Google as being equally as attractive to an academic job at a top-10 university. But, they can't always get them. Employers want to hire students who can write software that works and who can solve problems they didn't learn about in books. The perception is that newly minted PhDs in statistics often don't have those abilities.

Opinions were highly varied on how statisticians should react to the new demand for "data scientists." Some statistics departments are beginning to offer degrees in data science. The University of California at Berkeley has started a master's program in data science, a joint project of the computer science and statistics departments. The University of Warwick is offering the first undergraduate degree in data science in the UK, again with the collaboration of statistics and computer science.

Within a traditional statistics department, is there more that can be done to prepare students for a "data science" job? More computer science training seems to be a good idea, and it needs to go beyond simply learning more computer languages. The students need to learn how to produce software that is robust and timely. "It needs to finish running before we die," quipped Rafael Irizarry.

But the more time that students spend learning computer science, the less time they will have available for traditional training in statistics. The discussion of what parts of the "core" can be sacrificed, or if there even is a "core" that is fundamental for all students, produced even less agreement. A few voices tentatively called for less emphasis on the abstract mathematical foundations of the subject. However, some attendees felt that the unity of the subject was its strength, and they remembered fondly the days when they could go to a statistics meeting and understand any lecture. Even they acknowledged that things are changing; the trend is toward a field that is more diverse and fragmented. Should this trend be resisted or embraced? Will the pressure of Big Data be the straw that breaks the camel's back, or the catalyst that drives a long-needed change? On questions like these, there was nothing even approaching consensus.

Some participants in the meeting felt that the reward system at the postdoctoral level also needs to change. Currently, the promotion and tenure system rewards traditional publication in scientific journals. Generally speaking, the more theoretical journals are considered more prestigious than more applied journals, and statistics journals carry far more weight than journals in other disciplines, such as genomics or climate change. Virtually no statistics department would give as much weight to a piece of software as it would to a research paper. But if they are to prepare students for Big Data jobs, statistics departments will have to practice what they preach. They will have to reward faculty equally for theoretical research, applied contributions, and statistical programming. This would require a major cultural shift in many statistics departments.

Some worry that inertia and uncertainty will lead to an absence of a response, which would be the worst course of all. "I believe that the statistical sciences are at a crossroads, and that what we do currently … will have profound implications for the future state of our discipline," wrote Marie Davidian, past president of the American Statistical Association, in response to a pre-conference request for comments. "The advent of Big Data, data science, analytics, and the like requires that we as a discipline cannot sit idly by … but must be proactive in establishing both our role in and our response to the 'data revolution' and develop a unified set of principles that all academic units involved in research, training, and collaboration should be following. … There are many in our profession who are furious over 'data science' and the like and believe we should be actively trying to discredit these and work toward renaming anything having to do with data as 'statistics.' At this point, these new concepts and names are here to stay, and it is counterproductive to spend precious energy on trying to change this. We should be expending our energy instead to promote statistics as a discipline and to clarify its critical role in any data-related activity."

Terry Speed, winner of the 2013 Prime Minister's Prize for Science in Australia, offered a slightly different point of view in *Amstat News* after the meeting: "Are we doing such a bad job that we need to rename ourselves data scientists to capture the imagination of future students, collaborators, or clients? Are we so lacking in confidence … that we shiver in our shoes the moment a potential usurper appears on the scene? Or, has there really been a fundamental shift around us, so that our old clumsy ways of adapting and evolving are no longer adequate? ... I think we have a great tradition and a great future, both far longer than the concentration span of funding agencies, university faculties, and foundations. … We might miss out on the millions being lavished on data science right now, but that's no reason for us to stop trying to do the best we can at what we do best, something that is far wider and deeper than data science. As with mathematics more generally, we are in this business for the long term. Let's not lose our nerve." ❖

SECTION 3. Current Trends and Future Challenges in Statistics: Other Topics

3.1 The Reproducibility Crisis

One of the most controversial topics at the Future of Statistics workshop, after Big Data, was the problem of reproducibility in scientific research. While opinions vary as to how big the problem is, major science magazines and even the U.S. Congress have taken note.

In 2005, statistician John Ioannidis started the debate with a widely read and provocatively titled paper called "Why Most Published Research Findings Are False." His conclusion was based on a simple mathematical argument, combined with an understanding of the way that statistical significance tests, specifically p-values, are typically used. (See sidebar, "Reproducibility: Two Opposing Views.")

Though the details may be complex, a clear message emerges: A surprisingly large percent of claimed discoveries in the medical literature can be expected to be wrong *on statistical grounds alone*. Actual evidence based on the scientific merits suggests that the problem is even worse than advertised. Researchers from Amgen were able to replicate only 6 out of 53 supposedly classic results from cancer research. A team at Bayer HealthCare reported in *Nature* that when they tried to replicate 67 published studies, the published results were "completely in line with our in-house findings" only 14 times. These are all real failures of replication—not projected failures based on a statistical model. More recently, a network of nearly 100 researchers in social psychology attempted to replicate 27 studies. These were not obscure results; they were picked because of their large influence on the field. Nevertheless, 10 of the replication efforts failed completely and another five found smaller effects than the original studies.

These examples point to a larger systemic problem. Every scientific researcher faces pressure to publish articles, and those articles that report a positive new discovery have a great competitive advantage over articles with a negative result. Thus there is a strong selection bias in favor of positive results, even before a paper sees print.

Of course, the scientific method is supposed to weed out inaccurate results over time. Other researchers should try to perform the same experiment (this is called "replication"), and if they get different results, it should cast doubt on the original research. However, what should happen is not always what does happen. Replication is expensive and less prestigious than original research. Journals tend not to publish replication studies. Also, the original research paper may not contain enough information for another researcher to replicate it. The methodology may be described incompletely, or some data or computer programs might be missing.

In 2013, the journal *Nature* introduced new policies for authors of life-science articles, including an 18-point checklist designed to encourage researchers to be more open about experimental and statistical methods used in their papers. *PLoS One*, an open online journal, teamed with Science Exchange to launch a reproducibility initiative through which scientists can have their research validated (for a fee). "Reproducibility" has become a new catchword, with a subtle distinction from "replication." In the era of Big Data and expensive science, it isn't always possible to replicate an experiment. However, it is possible to post the data and the computer software used to analyze it online, so that others can verify the results.

The reproducibility problem goes far beyond statistics, of course, because it involves the entire reward structure of the scientific enterprise. Nevertheless, statistics is a very important ingredient in both the problem and the remedy. As Leek commented in a blog post, "The vast majority of data analysis is not performed by people properly trained to perform data analysis." Along with all the other measures described above, scientists need more access to qualified statisticians—which means that more statisticians need to be trained. Also, the statistical training of subject matter experts (scientists who are not statisticians) needs to be improved. The stakes are high, because the more the public reads about irreproducible studies, the more they will lose their trust in science.

3.2 Climate Change

One of the biggest stories in science in the 21st century, climate change, is also a field that is greatly in need of more statistical expertise. As Doug Nychka pointed out at the London workshop, climate is inherently a statistical concept. It is more than just the average temperature for a given day in a given place. It is a probability distribution over the entire range of weather possibilities. For New Orleans on August 29, the average high temperature is in the mid-80s, with 9 mph winds and a 40 percent chance of rain. Yet from the perspective of climate, it is surely just as important to realize that there may be 7 inches of rain and 125-mph winds on rare occasions. Those were the weather conditions on the day that Hurricane Katrina hit, in 2005.

Though climate studies and statistics seem to be natural allies, the connections between the two disciplines have not been as strong as one would expect. Climate models have tended to focus on average behavior, not extreme events. Yet the unusual events are the most damaging ones—hurricanes, droughts, floods. Statistics has a lot to say about forecasting extreme events. But you wouldn't know that from a 2012 report on weather extremes by the Intergovernmental Panel on Climate Change (IPCC). As Peter Guttorp pointed out in the *Annual Review of Statistics and Its Applications*, only four out of 750 listed authors of the IPCC report are statisticians, and "extreme value theory is mentioned in just one place in 542 pages of text."

The sources of uncertainty in climate forecasts are numerous. First, there are uncertainties in understanding the past climate—issues of modeling measurement error, interpolating sparse measurements taken at only a few places around the world, and reconciling measurements made with different equipment or methods.

Second, there are uncertainties in the parameters that enter into climate models. The models themselves are deterministic and based on the laws of physics, but they include quantities that scientists simply cannot measure—say, for instance, the heat capacity of the oceans. In the past, climate scientists have "tuned" the parameters of their models so that simulations of past climate match the historical record. This tuning has largely been undocumented and not performed in a statistically rigorous fashion. Inference of parameters from observed data should be a perfect application for Bayesian statistics, but it has not been done that way in practice.

Reproducibility: Two Opposing Views

Here are the basic arguments that John Ioannidis used to claim that more than half of all scientific discoveries are false, and that Leah Jager and Jeffrey Leek used to rebut his claim.

Most Published Discoveries Are False. Suppose that scientists test 1,000 hypotheses to see if they are true or false. Most scientific hypotheses are expected to be novel, perhaps even surprising, so *a priori* one might expect 90 percent of them to be false. (This number is very much open to debate.) Of the 900 false hypotheses, conventional statistical analysis, using a *p*-value of 5 percent, will result in 45 being declared true. Thus one would expect 45 "false positives." For the 100 true hypotheses, the probability of detecting a positive effect is called the "power" of the experiment. A typical target that medical researchers strive for is 80 percent. Thus, Ioannidis argued, 80 of the 100 true hypotheses will be declared true. These are "true positives." Both the false and true positives are presented as "discoveries" in the scientific literature. Thus, 45 out of 125 (36 percent) published discoveries will actually be false. If the *a priori* likelihood of a hypothesis being true is smaller, or if the power is lower, then the false discovery rate could be more than 50 percent, hence the title of Ioannidis' paper.

Most Published Discoveries Are True. Like Ioannidis' paper, Jager and Leek's estimate was not based on evaluating the actual scientific merit of any papers. However, it did go a step beyond Ioannidis in the sense that it was based on empirical data. They reasoned that for the false hypotheses, the *p*-value should simply be a random number between 0 and 1. On the other hand, for the true hypotheses, they argued that the *p*-values should follow a distribution skewed toward zero, called a beta-distribution. They collected all the *p*-values reported in every abstract published in five major medical journals over a 10-year span and computed which distribution best matched the *p*-values they found. The result was a mix of 14 percent false positives, 86 percent true positives. Hence, they concluded, about 86 percent of published discoveries in those five journals are true, with only 14 percent false.

Downscaling is a third source of uncertainty. Global climate models are conducted at a very coarse scale, and then the results are entered into finer-scale regional models to make forecasts on a country level. However, it would be better to transfer the whole distribution of possibilities, not merely a single number, from the large-scale model to the smaller one. This is called propagating the uncertainty.

Likewise, statisticians can help with what climate scientists call "detection and attribution": Is an observed change significant, and if so, what is causing it? This is a problem in hypothesis testing. For example, one question that is puzzling climate scientists now is the apparent slowdown in global warming from 1998 to 2012, which none of the climate models really predicted. (As Gabi Hegerl said at the workshop, the observations are "uncomfortably close to the edge of the distribution.") Is it a random deviation? Is there an exogenous cause (e.g., solar activity)? Or is there a flaw in the models? Some have suggested that the oceans are absorbing more heat than expected, which would mean that the models had an incorrect parameter.

The purpose of this long list of uncertainties is not to criticize climate science, but to indicate the many places where statisticians (the experts in uncertainty) can play a greater role. Some of these questions may need to be answered by experimenting with the climate models. Unfortunately, the full-scale computer models are so complicated that it is not really possible to run controlled experiments with hundreds of trial runs. Instead, it may be possible to simulate the simulations—in other words, to develop simpler "emulators" that predict reasonably accurately what the large-scale models will do. "The most important thing is to get the climate community to do designed experiments," said Nychka.

"The problem, as I perceive it, is not that climatologists are ignoring statisticians, but that there isn't a big enough community of statisticians who are willing or able to put their energies into addressing these problems," said Richard Smith at the workshop. Any statisticians who wish to help will be welcomed with open arms, Smith believes, as long as they stay focused on actual questions about the climate, rather than purely theoretical or methodological issues.

The overheated political climate that surrounds global warming has made it more difficult to talk openly about the uncertainties in climate models. Some climatologists may be reluctant to give their opponents an excuse to ignore their message. However, failing to take the uncertainty into account would be more damaging to the science in the long run. If climate modelers say that global average temperatures will increase by 3 degrees, they are almost certain to be wrong. If they say that the increase will be 1.5 to 4.5 degrees, they have a much better chance of being right. The climate skeptics may point to the lower end of the confidence interval and say that we have nothing to worry about. But, as Nychka points out, the actual amount of climate change is just as likely to be at the high end of the confidence interval. As the example of New Orleans during Hurricane Katrina shows us, it is important to be aware of all the possibilities.

3.3 Updating the Randomized Controlled Trial

Earlier in this report, the randomized controlled trial (RCT) was presented as one of the signature contributions of statistics to scientific research.

Nevertheless, some weaknesses of the RCT have become apparent—indeed, have been apparent for a long time. It's only natural for a 65-year-old innovation to become a little bit shopworn and to require an upgrade.

In this section, we will describe two recently tested experimental designs that are quite different from one another, but both address limitations of the conventional RCT design. The two stories, somewhat confusingly, share one word in common. The two designs are *adaptive clinical trials* and "SMART trials," which are designed to study *adaptive interventions*. To lessen the confusion, we will use another term that has appeared in the literature—"dynamic treatment allocation"—instead of "adaptive intervention."

An adaptive clinical trial is one whose experimental conduct can change as the trial progresses, provided that the changes are determined by a well-documented protocol. Typical adaptations include stopping a trial entirely if a treatment proves to be extremely successful or unsuccessful, or eliminating the less successful therapies in a multi-armed trial. It is worth noting that any such modifications have to be done in a way that is planned before the trial begins, otherwise the data may be subject to selection bias. A more sophisticated type of adaptive study can change incrementally. If a particular treatment is seen to be more successful than another one early in the trial, the probability of new patients being enrolled in the more promising therapy can be gradually increased. This makes sense both from a humanitarian and a scientific point of view. Alternatively, it may turn out that certain groups of patients respond to one treatment better than other groups do. If so, this information can also be used to guide the assignment of new patients. In effect, the trial can generate new hypotheses on the fly.

Unfortunately, such adaptations are difficult to incorporate into a traditional RCT, in which the experimenters are not allowed to see any data while the experiment is in progress. This precaution, called "blinding," has proved to be an important safeguard against experimenter bias. However, with careful advance planning, an adaptive trial can likewise be blinded. For example, the decisions can be made automatically by a computer program, programmed in advance with instructions on how to modify the trial under various contingencies. The computing algorithm is allowed to see the incoming data, but the experimenters do not.

The analysis of an adaptive trial is likewise more demanding than a traditional RCT. The probability distribution of the experimental outcomes under various hypotheses can't be taken out of a textbook; it has to be determined through thousands of computer simulations. The simulations allow the researchers to estimate, for example, the chances that one drug will go on a "lucky streak" and be assigned to more patients than it deserves. In the era before big computers, such large-scale simulations were unfeasible, and it made sense for scientists to use conceptually simpler RCT designs with no adaptation. But in the 21st century, there is no reason, except for tradition, to stick to classical methods.

> The problem, as I perceive it, is not that climatologists are ignoring statisticians, but that there isn't a big enough community of statisticians who are willing or able to put their energies into addressing these problems…

In the first large-scale adaptive trial, called I-SPY 2, Don Berry and co-investigator Laura Esserman simultaneously studied the effectiveness of five breast cancer drugs made by different manufacturers on 10 subpopulations of cancer patients. By the end of 2013, they had identified two drugs with a high probability of success in a Phase III study. One drug, called veliparib, was effective against triple-negative breast cancer; the other, called neratinib, showed promise against HER2-positive, estrogen-negative, and progesterone-negative cancers.

"Adaptive design is like driving a car with your eyes open," wrote Berry in 2010. "However, the driver is an automaton that, for example, is programmed to turn around when coming to what the results of the trial indicate is a dead end."

A different set of issues is addressed by dynamic treatment allocation. In many drug studies that deal with an acute illness, the treatment consists of the administration of one drug for a defined length of time. Such treatments are well suited for analysis with a classic RCT design. However, in chronic illnesses or behavioral disorders such as alcoholism or ADD, a clinical intervention takes place over a much longer period and may involve several steps. Also, the patient population for mental disorders is quite heterogeneous. Some treatments may work for some patients and not for others. An intervention that works in one patient for a while may stop working. Thus, researchers on such disorders are interested in questions like these: What treatment should I try

Heterogeneity → Sequences of Treatment
High heterogeneity in response to any one treatment
- What works for one child may not work for another
- What works now for a child may not work later
In other... relapse rates are high, lack of adherence... burden may be common
ROYAL STATISTICAL SOCIETY

Susan Murphy speaking about statistics in autism research.

first? How can I decide if the treatment is not working? If it is not working, what is the best option to try next—perhaps increase the intensity of that treatment, or try a completely different one?

These questions call for a different study design, called a SMART design. (SMART stands for Sequential Multiple Assignment Randomized Trial.) As Susan Murphy explained at the workshop, a SMART is similar to a factorial design (an experiment with two or more treatments, each with two or more dosage levels), but it involves an additional time component. This allows the researcher to answer some new kinds of questions. He or she can investigate not only whether intervention A is better than intervention B, but also whether A followed by B might be better than both of them.

A conventional RCT investigates only one "decision point" in isolation and does not consider formally what might happen next in a patient's treatment. By contrast, a SMART is a clinical trial with more than one decision point and several options at each. At each decision point, the patients are re-randomized. The options at each decision point have to be clearly stated in advance. The decisions are based on observed variables, and the decision rules (which may be deterministic or probabilistic) must also be prescribed in advance.

At present, SMART studies have been conducted or are in progress for treatments of autism, ADHD, drug and alcohol abuse, and depression. Murphy, a MacArthur Foundation Fellow, reported on a study of autistic patients that she had consulted on that showed (counter to expectations) that giving autistic patients an iPad over a period of several weeks improved their ability to form spontaneous utterances.

Because adaptive trials and dynamic treatment allocation sound so similar (even statisticians get confused), it is worth highlighting the differences. First, adaptive trials are an experimental design, which is a statistical concept. Dynamic treatment allocation is a medical (not statistical) concept, which calls for a different experimental design (a SMART). In adaptive trials, the trial itself evolves: The decision rules change and the treatment of later patients is affected by the results of earlier patients. In SMARTs, the decision rules are static and each individual patient's treatment evolves. The adaptive decisions depend on the patient's own previous history, not the history of other patients.

In spite of their differences, one comment seems applicable to both of these innovations. Whenever statisticians introduce a new method, such as a new experimental design, they will have to make a convincing case to their collaborators, to funding agencies, and to journals. In medicine, lives and money are at stake, so all of the above stakeholders tend to be conservative. To overcome this natural caution, Murphy and Berry both emphasized the importance of learning to communicate with the medical experts *in their own language*.

3.4 Statistics versus Conventional Wisdom

While "big science" and Big Data" tend to garner more attention, the Workshop on the Future of Statistical Sciences also showcased some less widely known applications. Two of these, having to do with dietary science and demography, are discussed here.

First, Sue Krebs-Smith and Ray Carroll reported on a new method of assessing national dietary patterns, developed at the National Cancer Institute, called the Healthy Eating Index-2010 (HEI). The motivation for coming up with a standard metric of diet quality is self-evident: Diet plays a significant role in many diseases, such as hypertension, diabetes, cancer, and osteoporosis. It has been estimated that the United States spent $147 billion in 2008 on medical care for obesity-related diseases. To answer questions such as "What are we doing?" and "What could we do better?" we need to compare our diets with an objective standard.

However, it's hard to sum up the quality of our diet in a single number, because there are many components to a healthy diet. The HEI scores diets on 12 components (e.g., total fruit and total vegetables (for these, a higher intake is better), sodium, and empty calories (for which a smaller intake is considered better). The quantity of calories consumed is also an important factor, but it is not a category in the HEI. Instead, each of the variables is evaluated relative to the total number of calories consumed (e.g., cups of green vegetables per 1,000 calories). This makes it possible not only to compare individuals, but also to compare producers, markets, restaurants, and other points along the food supply chain.

The HEI poses some interesting challenges to a statistician. For example, the 12 categories in the HEI are all converted to a numerical rating (say, a number from 0 to 5). The distribution of scores in most of the categories is highly skewed. For six of the categories, zero is the most common response, a situation that is common enough in statistics that there is a name for it (a "zero-inflated" probability distribution).

Another serious issue is measurement error. Nutritionists would really like to understand the usual, everyday intake of individuals. Unfortunately, it is not possible to follow someone and observe every meal (and midnight snack) for weeks on end. Nor can most people accurately answer a question about the average amount of green vegetables they consume in a day. Thus, nutritionists typically rely on interviewers probing a person about their intake just in the last 24 hours. Most people can answer such questions in great detail. The problem is that a person's last 24 hours may not be typical. The intake of some foods, such as fish, varies a great deal from day to day and is often zero. This adds additional uncertainty (within-person variation) to the data, and it also accounts for the tendency of certain scores to cluster around zero.

In a rare "good news" moment about American eating patterns, Carroll showed that the biases introduced by the 24-hour recall studies had led nutritionists to an overly pessimistic conclusion. It appeared that 25 percent of children in the United States had HEI scores of 40 or below, a score that Krebs-Smith once called "alarmingly bad" from a nutritional point of view. However, when Carroll statistically corrected the HEI scores for measurement bias and zero-inflation effects, he found that only 8 percent of youngsters had such an "alarmingly bad" diet.

Future projects for the National Cancer Institute include correlating HEI scores to health outcomes and cluster analysis to determine if identifiable subgroups of the population have consistent problems with their diet. One exciting prospect on the horizon is the use of cell phones to replace interviewer-administered questionnaires. Volunteers can now be asked to photograph every meal they eat, not just for one day, but for a whole month. A computer program can then convert the photographs to HEI scores. This has the potential to improve the quality of data compared to the 24-hour recall questionnaires. Over the long term, it's conceivable that nutritionists will be able to follow the changes in a person's diet and the health consequences from cradle to grave.

Moderator Sastry Pantula looks on while Sue Krebs-Smith responds to questions about her presentation on the use of statistics in evaluating dietary patterns

Participants discuss one of the many presentations at the London workshop.

In another example from the London workshop, Adrian Raftery shared his experience working on world population for the United Nations. Every

Why Bias Matters

Abraham Wald, an Austrian statistician, fled his native country in 1938 and subsequently contributed to the U.S. effort in World War II as a member of the Statistical Research Group (which eventually became part of the Center for Naval Analyses). One of his projects was to estimate the survival rates of aircraft that had been shot in various locations. The data Wald was given showed, for instance, that returning aircraft were riddled with bullet holes to the wings but the engines had emerged relatively unscathed. Some people might conclude, therefore, that the wings needed more reinforcement.

Wald realized that the exact opposite was the case, because he thought about the source of his data. The reason that the returning planes had few bullet holes in the engine was that the planes that *had* been hit there had crashed. In other words, there was a strong *sampling/selection bias*. The data did not come from a random sample of all the airplanes that had flown, but only the planes that had survived. Those planes by definition had less serious, non-fatal damage. Thus, he argued, the most important place to reinforce the hull was precisely the engine.

The above account (a Wikipedia version) considerably oversimplifies what Wald actually did. The data were not so black and white; about half of the planes shot in the engine area did make it back somehow. He constructed a statistical model that allowed him to infer the most likely survival rates of planes based on the number and location of hits, in effect filling in the missing data on the planes that had crashed.

Nevertheless, the "Wikipedia version" of the story illustrates very clearly the value of a good statistician. It also shows that if the data are collected in a biased manner, the story they appear to be telling may be diametrically opposed to the story they are actually telling.

two years, the UN publishes a projection of future population trends in every country of the world. Governments and social planners around the world use this information to, say, plan the number of schools that need to be built or to project how many retirees will need to be supported by government pensions.

In the first decade of this century, projections had shown that the world's population might level off at about 9 billion people by mid-century. However, the new forecast by the United Nations Population Division, using statistical methods developed by Raftery and his colleagues (including Patrick Gerland of the United Nations) shows that the supposed leveling-off is in fact unlikely to happen. Raftery's research forecasts a population of 9 to 13 billion by the year 2100. This is the first time the world's population has been forecasted by a major agency using a statistical method that explicitly incorporates uncertainty.

The UN's previous method, used in its publication *World Population Prospect*s through 2008, is completely deterministic. It gathers data on the current population of each country by age and sex and on fertility, migration, and mortality rates. It then projects future country-wide trends in total fertility rate and life expectancy, using empirically observed functions that agree with historical data. Most notably, the fertility rate in all countries has followed a broadly similar pattern. Prior to industrial development, fertility rates are very high (4 to 8 children per woman). As the nation develops, it goes through a "fertility transition," a rapid decrease in the fertility to below the replacement rate of 2.1 children per woman. Finally, in certain countries (mostly in Europe), a third phase of gradual recovery toward the replacement rate has been observed. (It remains controversial whether this same pattern will hold for all countries.) After these historical trends, adapted to each country, have been used to project the total life expectancy and mortality rate, another deterministic calculation converts these overall rates to distinct rates within each five-year age cohort.

As this report has already pointed out several times, a prediction without uncertainty estimates has a nearly 100 percent chance of being wrong. As a nod to uncertainty, the UN publication also includes "high" and "low" estimates for each country. However, these are not based on any sort of probabilistic method. Instead, the UN simply re-computes the deterministic forecast, using fertility rates one-half child higher (for the high estimate) and one-half child lower (for the low estimate). As Raftery has shown, this non-probabilistic method tends to overstate the uncertainty for countries that have completed the fertility transition, while understating the uncertainty for countries that are near the beginning of the transition. UN demographers realized that the existing method was not adequate and invited Raftery to help them develop a probabilistic alternative.

The new method is based on Bayesian hierarchical models for fertility and mortality. The idea is to treat the total fertility rate and life expectancy as random variables that have a certain probability of going up in a given year and a certain probability of going down. For a country going through the fertility transition, the overall trend in the fertility rate is down, but there can be upward or downward blips due to random events such as war, famine, or economic booms or busts. The model is called "hierarchical" because the forecast for one country uses past information from other countries. The model is estimated using Markov chain Monte Carlo methods (see §1.3).

There are two reasons that the new estimates came out with a higher population than before. The first is that the fertility rate in Africa is not decreasing as fast as had been assumed based on the experiences in Asia and Latin America in the past 60 years. The second is the use of the Bayesian hierarchical model, which incorporates recent data more directly and fully than the previous deterministic, assumption-driven approach. The United Nations issued its first fully probabilistic projections, using Raftery's method, on an experimental basis in 2012. It is considering the adoption of the probabilistic method for its official projections beginning in 2014.

Both of the above case studies illustrate how effective statistical reasoning can be for persuading subject matter experts to reconsider the conventional wisdom in their fields. In the first example, nutritionists had not properly appreciated the way that their data gathering method (the 24-hour recall questionnaires) had biased their statistics. The conventional wisdom turned out to be too pessimistic. In the second example, demographers did not realize how their deterministic projections had misled them about the amount of uncertainty. In this case, the conventional wisdom turned out to be too optimistic. ❖

SECTION 4.
Conclusion

The Workshop on the Future of Statistics did not end with a formal statement of conclusions or recommendations. However, the following unofficial observations may suffice:

1. The analysis of data using statistical methods is of fundamental importance to society. It underpins science, guides business decisions, and enables public officials to do their jobs.
2. All data come with some amount of uncertainty, and the proper interpretation of data in the context of uncertainty is by no means easy or routine. This is one of the most important services that statisticians provide to society.
3. Society is acquiring data at an unprecedented and ever-increasing rate. Statisticians should be involved in the analysis of these data.
4. Statisticians should be cognizant of the threats to privacy and confidentiality that Big Data pose. It will remain a challenging problem to balance the social benefits of improved information with the potential costs to individual privacy.
5. Data are coming in new and untraditional forms, such as images and networks. Continuing evolution of statistical methods will be required to handle these new types of data.
6. Statisticians need to re-evaluate the training of students and the reward system within their own profession to make sure that these are still functioning appropriately in a changing world.
7. In particular, statisticians are grappling with the question of what a “data scientist” is, whether it is different from a statistician, and how to ensure that data scientists don’t have to “reinvent the wheel” when they confront issues of uncertainty and data quality.
8. In a world where the public still has many misperceptions about statistics, risk, and uncertainty, communication is an important part of statisticians’ jobs. Creative solutions to data visualization and mass communication can go a long way.

We conclude with some observations on statistical education, which was a major topic of discussion at the London workshop, even though there were no formal lectures about it.

In the United States, the number of statistics degrees and the enrollment in introductory statistics courses is increasing robustly. The numbers of bachelor’s and master’s degrees awarded in statistics have both roughly doubled in the last 10 years. The representation of women in statistics programs is much better than it is in comparable disciplines such as mathematics, physics, and engineering. At the undergraduate level, enrollment in introductory statistics courses has gone up by 90 percent from 1995 to 2010.

Clearly, some students are getting the message that statistics is a useful major, and many of them are undoubtedly attracted by the job possibilities. However, statistics departments need to do a better job of preparing them for the jobs that are actually available and not necessarily to become carbon copies of the professors. Some suggestions include the following:

- *Working on communication skills.* Statisticians have a deep understanding and familiarity with the concept of uncertainty that many other scientists lack. They will only be able to disseminate their knowledge of this critical concept if they can convey it readily and with ease.
- *Working on team projects, especially with non-statisticians*. The workshop itself modeled this behavior, as most of the speakers who were statisticians were paired with a non-statistician who is an expert in the subject-matter area under discussion. In most cases, the two speakers were collaborators.
- *Training on leadership skills.* There was a strong sentiment among some workshop participants that statisticians are pigeonholed as people who support the research of others, rather than coming up with original ideas themselves.
- *Strong training in an application field.* This again may help prepare the students to steer the direction of research, rather than following it.
- *More exposure to real "live" data.* Many students will learn best if they can see the applicability to real-world problems.
- *More exposure to Big Data, or at least reasonably Big Data that cannot be analyzed using traditional statistical methods or on a single computer.* Students need to be prepared for the world that they will be entering, and Big Data seems to be here to stay.
- *More emphasis on computer algorithms, simulation, etc.* To prepare for engineering-type jobs, students need to learn to think like engineers.

> I am still amazed by the power of statistics. … Because of statistics, we are able to have a glimpse of the future, to understand …

At the high-school level in the United States, there is also good news and bad news. The Advanced Placement course in statistics has grown rapidly in popularity and increased the awareness of statistics in high schools. There is absolutely no reason why high-school students cannot have rewarding experiences in statistics. The relevance of statistics to real life is much more readily apparent than most other high-school math courses. However, as Richard De Veaux observed, "Statistics education remains mired in the 20th (some would say the 19th) century." As Big Data changes statistics, statistical education will also have to change.

As for the situation of statistics education outside of North America and Europe, there was not enough expertise present at the London workshop to draw any conclusions. However, we will end the report with one anecdotal comment.

Statistics is a vital component to cancer research. High school student Shannon Hunt illustrated this artistically in her photo, the winning entry in the Statistics2013 Photo Contest.

In advance of the workshop, the organizing committee invited written comments not only from people who had been invited to attend the workshop, but also from anyone else who wanted to send one. One non-participant comment came from a student named Nilrey Cornites, who had just completed his undergraduate degree in statistics in the Philippines. This student's (mildly edited) response could serve as an inspiration to all statisticians.

"During my elementary and secondary education, I never knew anything about statistics. … In my first year in college, I was enrolled in BS mathematics and there I appreciated the beauty of statistics in our introduction to statistics class. I was so in love with statistics that I shifted to statistics as my major.

"I am still amazed by the power of statistics. … Because of statistics, we are able to have a glimpse of the future, to understand … the significant effect of a new product or medicine, and to understand the weather. Statistics saves lives, lowers the cost, helps ensure success, and improves things and processes.

"In the Philippines, most of the people and even companies and government or non-government agencies don't yet appreciate statistics as a very powerful and reliable quantitative tool to help them in their decisionmaking, because they don't even know what is statistics and its function.

"And as a young dreamer to become a full-fledged statistician in the future, I know that in the very near future they will seek after me, seek after us (statisticians) asking for our assistance. … *We are the future tellers* [emphasis added] and someday they will flock to us to see their future. … I am a very proud young statistician."

This student has laid out a very ambitious goal and a daunting responsibility for statisticians. If statisticians are to be the future tellers, they must stay humble, always presenting uncertainty as an integral part of their forecasts. They need to be willing to tell unpopular truths. They need to be ingenious and innovative in taking advantage of new technology and new sources of information. All of these things are part of the history of statistics, and, with any luck, they will also be part of its future. ❖

ACKNOWLEDGEMENTS

There are many people responsible for this document and all that went into its creation. We will mention a few, but we hope all who were involved know that their efforts were appreciated and see their contributions in this report.

We thank all the sponsors of the London workshop (listed elsewhere in this document). Special thanks go to the Board of Directors of the American Statistical Association, who pledged major support sufficient to ensure that the workshop would happen, and to the Royal Statistical Society, for graciously and effectively hosting the workshop.

We thank our fellow members of the Organizing Committee (listed below) for their hard work in creating the format for the workshop and for lining up the participants. Of course, we also thank the participants for their thoughtful contributions, both to the workshop itself and to the various drafts of this report. Special thanks to Sastry Pantula for many contributions to workshop development and for serving as moderator.

Science writer Dana Mackenzie did the lion's share of the writing of this report, and we thank him for his professionalism, his clear prose, and his keen ear. Workshop participant Tom Louis went well beyond the call of duty in providing editing assistance. Thank you, Tom. Many thanks also to ASA staff for final editing and formatting of the report.

The participants in the workshop are representative of a deep and diverse set of statistical colleagues around the globe who work diligently to further the discipline and profession of statistics. We thank this worldwide community of statisticians for their work in advancing science, informing policy, and improving human welfare. Without them and those they are preparing, there would be no future of the statistical sciences.

Sincerely,
David Madigan (Columbia University) and Ron Wasserstein (American Statistical Association)
Workshop organizers

The Organizing Committee of the London workshop also included:

- Peter Bartlett, University of California, Berkeley
- Peter Bühlmann, ETH Zurich
- Raymond Carroll, Texas A&M University
- Susan Murphy, University of Michigan
- Gareth Roberts, University of Warwick
- Marian Scott, University of Glasgow
- Simon Távare, Cancer Research United Kingdom Cambridge Institute
- Chris Triggs, University of Auckland
- Jane-Ling Wang, University of California, Davis
- Khangelani Zuma, HSRC South Africa

STATISTICS2013.ORG

INTERNATIONAL YEAR OF
STATISTICS