STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 149–158, DOI 10.21307/stattrans-2018-009

SOME RESULTS FROM THE 2013 INTERNATIONAL YEAR OF STATISTICS

Jan Kordos[1]

ABSTRACT

There are presented in this report, seven case studies of the uses of statistics in the past and present. I do not intend these examples to be exhaustive. I intend them primarily as educational examples for readers who would like to know: *What is statistics good for?* Also, to encourage the readers to study detailed reports from the 13 International Year of Statistics given in the notes of this report.

Key words: statistics, International Year of Statistics, Bayesian statistics, frequentist, data quality, official statistics, science of uncertainty, Markov Chain Monte Carlo (MCMC), Big Data.

1. Introduction

In 2013, six professional societies[2] declared an International Year of Statistics to celebrate the multifaceted role of statistics in contemporary society:

a) to raise public awareness of statistics, and;

b) to promote thinking about the future of the discipline.

In addition to these six societies, more than 2,300 organizations from 128 countries participated in 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. The organizers made the freely available lectures and discussions at Internet[3].

In Poland several organizations, societies and universities participated in the celebration. The Central Statistical Office of Poland and the Polish Statistical Association organized on 17-18 October 2013 a scientific conference entitled *Statistics – Knowledge – Development*[4].

[1] Warsaw Management University. E-mail: jan1kor2@gmail.com.

[2] 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

[3] Statistics and Science – A Report of the London Workshop on the Future of the Statistical Sciences. http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf

[4] Some papers have been published in Statistics in Transition new series.

The Warsaw Management University and the Polish Statistical Association organized on 25-26 November 2013 a scientific conference entitled *Statistics in Service of Business and Social Sciences*[5].

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 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.

2. Purpose of this report

This report is projected primarily for people who are not experts in statistics. It is intended as a resource:

a) for students who might be interested in studying statistics and would like to know something about the field and where it is going;
b) for policymakers who would like to understand the value that statistics offers to society, and;
c) for people in the general public who would like to learn more about this often misunderstood field.

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[6], 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.

[5] E., Frączak, A. Kamińska, J., Kordos (Eds), (2014). Statistics – Business and Social Sciences Applications (in Polish). Available at: http://www.kaweczynska.pl/wydawnictwo/publikacje/wazniejsze-publikacje.

[6] CEOs communicate, collaborate, and exchange information on Earth observation activities, spurring useful partnerships such as the Integrated Global Observing Strategy (IGOS), http://ceos.org/about-ceos/overview/.

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.

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.

3. Seven case studies of past "success stories" in statistics continued to the present day.

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:

1) 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.
2) The application of "***Bayesian statistics***" to image processing, object recognition, speech recognition, and even mundane applications such as spellchecking.
3) The explosive spread of "***Markov chain Monte Carlo***" methods, used in statistical physics, population modelling, 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").

4) ***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.
5) The discovery through statistical methods of "***biomarkers***"[7] – genes that confer an increased or decreased risk of certain kinds of cancer.
6) A method called "***kriging***[8]", which enables scientists to interpolate a smooth distribution of some quantity of interest from sparse measurements. Application fields include mining, meteorology, agriculture, and astronomy.
7) The rise of "***analytics***" in sports and politics in recent years. 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.

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.

The original algorithm was designed in the late 1940s by Nicholas Metropolis, Stanislaw Ulam, the Polish statistician, 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.

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

[7] The term "biomarker", a portmanteau of "biological marker", refers to a broad subcategory of medical signs – that is, objective indications of medical state observed from outside the patient – which can be measured accurately and reproducibly.

[8] kriging – optimal interpolation based on regression against observed z values of surrounding data points, weighted according to spatial covariance values, http://www.kriging.com/whatiskriging.html.

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.

4. BIG DATA

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 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:

a) Are the data representative?
b) What is the nature of the uncertainty?
c) 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.

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), which dwarf in size any databases statisticians encountered in the past. *Big Data* is a challenge for several reasons:

1) *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.

2) *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.
3) *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 are not) is guaranteed to increase, unless specific precautions are taken.
4) *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.
5) *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 different meetings, and indeed there was a minority sentiment that it is an overhyped topic that will eventually fade. Other topics that were discussed include:

i. *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.
ii. *Updates to the randomized controlled trial.* The traditional RCT[9] is expensive and lacks flexibility. "Adaptive designs[10]" and "SMART trials[11]" 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.
iii. *Statistics of climate change[12].* 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.
iv. *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.
v. *Communication and visualization.* The Internet and multimedia give statisticians new opportunities to take their work directly to the public.

[9] RCT (Randomized Control Trial) is a type of scientific (often medical) experiment which aims to reduce bias when testing a new treatment.
[10] http://adaptivedesigns.com/about.
[11] SMART-trials – a next generation platform intended for data acquisition in medical research and clinical trials, https://www.cognizant.com/SmartTrials.
[12] http://data.worldbank.org/topic/climate-change.

vi. *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.

vii. *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.

5. 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.

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.

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.
- 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.

7. Some conclusions

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 reevaluate 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 do not 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.

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.

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