The Impact of the New Astrostatistics on the Development and Future of Statistics

Joseph M Hilbe
Arizona State University and Jet Propulsion Laboratory

Presented at the
2013 ISI World Statistics Congress,
Hong Kong SAR, China
IPS057

Prior to considering the impact of astrostatistics on the future of statistics, it is reasonable to first surmise what we belief statistics to be like at some future date. This is not an easy task, since the manner in which analysts "do" statistics at any given time largely reflects the widespread computational capabilities available at the time. I will offer a suggestion as to what that capability will be in, let's suppose, 32 years (2045), and then reflect on how astrostatistics as a discipline can help bring that capability to fruition.

When I was an undergraduate in the mid 1960s, use of the university computer entailed submitting directions to the mainframe from a remote site in the library. Some departments also had an electronic connection to the central computer or computers, but typically these were limited to those in the physical sciences. Constructing code for execution of a statistical procedure on the mainframe was usually done using Fortran, but the statistical results would usually not be available until the next day. Making interactive amendments to the code was simply not possible. Statistical modeling was limited to what we now consider to be rather simple operations, although many new discoveries were made about the fundamentals of the subject.

It was not to the advent of Personal Computers in 1981 that interactive statistical analysis could take place on a widespread basis. Soon most academic departments had PCs, and academic priced computers were made available to faculty for personal purchase. Even with the ability to make immediate amendments to code in order to effect a better understanding of the data, the early PCs and mini-frames were still severely limited in terms of disk space and in particular RAM and speed of operations. Iterative methods -- such as estimation of non-linear models, sometimes took overnight or more to converge. MCMC type simulation was not even considered, except for the most basic models.

Issues of Big Data and analysis became important in some areas fairly soon after PCs gained a foothold in economics and health care. The first truly major Big Data project of which I am aware began in 1990. The United States Health Care Financing Administration (HCFA), which regulates the national Medicare program - a nationalized health care system for seniors 65 and older, and for those who are deemed disabled - constructed a data set called MedPar which contained 115 data fields for every Medicare patient hospital stay throughout the year in the U.S, beginning with 1990. The project, called the Medicare Infrastructure Project, of which I was lead statistician, was for four years, 1990-1993, and resulted in the analysis of some 17 million patient records per year. The data management task was to find a way to get the data, stored on mainframe computers at HCFA, into a format so that it could be evaluated by analysts working

on PCs. Analysts sought evidence of unusual activity that would help identify issues related to 'cost and care'. Given the limitations of RAM, and processing speed, the project was a success, and useful information to the fields of health outcomes, health economics, and data processing.

The solution for statistically analyzing the large MedPar data set in the early to mid 1990s was to break it into State components, or into observations stratified by diagnostic group (DRGs). Analysis could not be accomplished on the data as a whole. Innovative ways of abstracting needed observations out of the national data for various types of analyses were developed and were made known to Medicare analysts throughout the nation. Reporting was done using the same methodology.

I should mention that the German Health Reform data, commencing in 1984, was established in response to legislation in 1988 that initiated payment reforms aimed at reducing the frequency of patient visits to their physicians and to the length of hospital length of stay. The data was collected and analyzed to determine if the reforms were successful. The data was evaluated descriptively using mainframe procedures until the 1990's, about the time of MedPar, which seems to have stimulated Big Data analysis in various domains.

Look at the computing power we have 32 years after the first PC came on the market, and some 23 years after the initiation of MedPar. Even relatively inexpensive PCs are substantially more powerful than mainframe computers of that era - much less the PCs of the time. It is now perfectly feasible to store a year's MedPar data on a CD, or in a single file on a PC. In the early 1990s I stored MedPar data on multiple external drives -- six as I recall -- and developed software to read and append data on specified elements from the drives. Now such data can be read and evaluated from a single file. Regressions can be run on huge data having a millions or more observations, with the results being displayed within moments. Using Stata software on my laptop PC, I just ran a million observation regression of a continuous variable on a binary predictor. It took 0.07 seconds to calculate and display on the screen. I could only have imagined such a thing 20 years ago - much less 32 years ago.

Big Data has come to be regarded as data which is so large that it cannot be analyzed as a single data set using standard data management and analysis methods. I suspect that this will be the focus of data management technology in the next third of a century. Developing statistical methods to handle Big Data situations will also be a foremost focus of statistical analysis. This will entail enhanced parallel processing methods as well as thoroughly new generations of processing chips that are based on developments in nanotechnology.

Drivers of Big Data are currently those in the information science industry such as Google and Amazon. Huge amounts of advertizing money is at stake -- which motivates advances in technology to meet the desired needs. With respect to the development of enhanced statistical methods, I believe that environmetrics and astrostatistics/astroinformatics will lead the way.
Environmental statistics, which covers such areas as weather, climate change, oceanography, ecology, fisheries, forestry, water and soil conservation, and even epidemiology has a near equal interest in dealing with Big Data problems as do astronomers. Both handle huge amounts of data. And now that the large data sets are being built, both disciplines must find a way to best evaluate it.

Given the papers presented at the Statistical Challenges V conference mentioned earlier, and the invited and special topics session papers presented at the Dublin World Statistics Congress in 2011, it is clear that researchers in astronomy will be accessing data in the future that dwarfs current Big Data repositories such as the ongoing Sloan Digital Sky Survey which began in

2000. SDSS spawned a variety of ancillary projects, most of which have substantially enhanced our understanding of astronomical events.

Until recently many astronomers used statistical methods when analyzing astronomical entities and events, but the methods were not very sophisticated -- not up to par with the advances occurring in professional statistics. Of course, there are those who have been keeping up with statistical advances. With the recent establishment of the International Astrostatistics Association (IAA) and the astrostatistics/astroinformatics working groups authorized by the IAU and AAS, substantial advances in statistical methods for evaluating large scale data sets as well as how to deal with the other astrophysical issues are occurring

Astronomers have had experience in handling SDSS, and are actively preparing for managing and evaluating the truly massive amount of data that will be coming from LSST beginning in about 2022. They have a specific data set with which to work, and teams of astronomers - and statisticians - are already planning how to develop better statistical and data management tools in order to abstract the most information possible form the data. The LSST will be gathering not just petabytes, but multiple exobytes of data once it is in operation.

I cannot foretell what types of statistical methods will be in operation in the year 2043. I know that I will not be witness to them. But I suspect that new statistical methods combining appropriate aspects of both frequentist and Bayesian methods of analysis will be developed, and that data management will involve data storage capabilities which we do not have at present.

Big data problems will likely be at the forefront of astrostatistical concern in the future, as will it be of importance to the world of analytics in general. Other areas will also be of concern, such as signal detection of difficult to observe objects or events. Also, enhanced computing power will also allow for exact statistics to be used in place of traditional asymptotic methods for appropriate models. This latter promise for the mid 21 century is not of much interest in astrostatistics, at least at this time. It is doubtful that it will be in the future.

Astrostatistics has the promise to lead the way to the handling of huge data situations as the mid 21st century approaches. It is now possible to implement complex hierarchical Bayesian models, involved simulations, and interactive 3D graphing. We are in an electronic age that could not be imagined but by a few some 32 years ago..I suspect that in another 32 years there will be computing capabilities that are difficult to imagine except by those who are in the field of information and computing science, nanotechnology, and perhaps also by those who are at the cutting edge of environmetrics and astroinformatics. Computers may allow estimation of classes of dynamic nonlinear spatio-temporal models can provide more insight into data than can now be envisioned. The coordinated work of those in informatics, astrophysics, and statistical methodology will likely lead the way in dealing with Big Data. And it is this concern which I think will characterize much of future statistics until mid century.