International Journal of Movement Education and Social Science
Volume 7, Special Issue 03, 2018 www.ijmess.org
ISSN (Print): 2278-0793
ISSN (Online):2321-3779

Statistics and Data Science-Trend and Challenges in 21st Century

Deepak Kumar Routray

Designation– Head, Dept. of Statistics.,Ispat Autonomous College,Sector-16, Rourkela-3, Odisha

ABSTRACT

The basic ideas of statistics and data science are becoming almost a core competency for citizenship in the 21st Century. We can all see that the world is becoming more and more data-oriented, and statistics is becoming one of the most important pillars of our progress. With statistics, we test hypotheses and break myths. This paper encourages a big-tent view of data analysis. We examine how evolving approaches to modern data analysis relate to the existing discipline of statistics (e.g. exploratory analysis, machine learning, reproducibility, computation, communication and the role of theory). Finally, we discuss what these trends mean for the future of statistics by highlighting promising directions for communication, education and research.Data science is being essentially the systematic study of the extraction of knowledge from data. But analyzing data is something people have been doing with statistics and related methods for a while. Why then do we need a new term like data science when we have had statistics for centuries? Taking all these facts this paper reviews some ingredients of the current\Data Science moments, including the trend and about how/whether Data Science is correlated with Statistics and what are the trend and challenges in 21st century.

Key word**:** *Classical statistics, Computation, Data science, Field-by-Field, Hypothesis, 21st Century.*

I.INTRODUCTION:

Statistics is the science of making inferences and decisions under uncertainty; it is becoming increasingly relevant in the modern world due to the widespread availability of and access to unprecedented amounts of data and computational resources. Unlike classical Statistics, the need to process and manage massive amounts of data has become a key feature of modern Statistics. This aspect of managing and processing data is popularly referred to as "data science." Data Science is the competency to make sense of, and find useful patterns within data to better support decision making. Where Statistics is concerned with designing experiments and other data collection, Summarizing information to aid understanding, drawing conclusions from data and to estimate the present or predicting the future. Data science is the heartbeat of 21st century global economies, and innovations in sciences, engineering, business, and education which are becoming increasingly computationally and data-enabled.

"Data science has become a fourth approach to scientific discovery, in addition to experimentation, modelling, and computation," said Provost Martha Pollack."

"The web site for DSI gives us an idea what Data Science is: This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of

ISSN (Print): 2278-0793
ISSN (Online):2321-3779

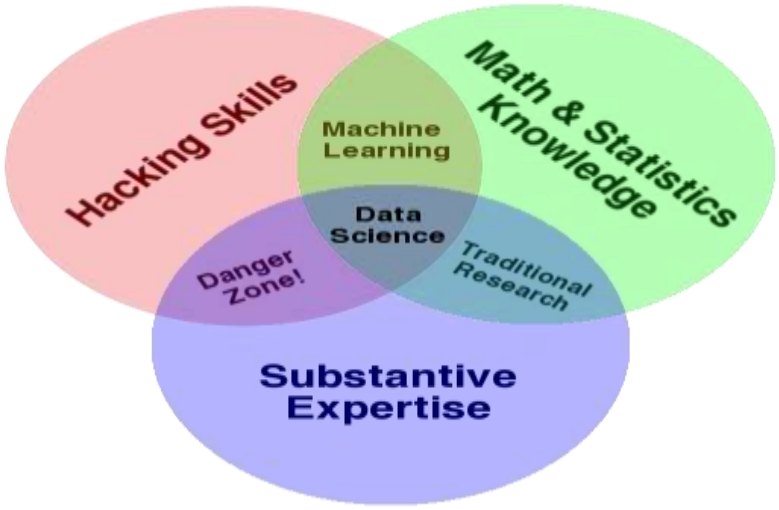
heterogeneous data associated with a diverse array of scientific, translational, and inter- disciplinary applications."

A simple definition of a data scientist is someone who uses data to solve problems. In the past few years this term has caused a lot of confusion in statistical industry, One might argue that data science is simply a rebranding of statistics (e.g. "data science is statistics on a Mac"4) but this comment misses the point. While data analysis has been around for a long time, its economics (costs of and value derived) have changed primarily due to technology driving the availability of data, computational capabilities and ease of communication. The most obvious advances are in computer hardware (e.g. faster CPUs, smaller microchips, GPUs, distributed computing). Similarly, algorithmic advances play a big role in making computation faster and cheaper (e.g. optimization and computational linear algebra).

There are many new/improving technologies which allow us to gather data in new, faster and cheaper ways, including: drones medical imaging, sensors (e.g. Lidar), better robotics, Amazon's Mechanical Turk, wear cables, etc. Finally, improved software (e.g. see Section 4.2.1) makes it faster, cheaper and easier to communicate the results of data analysis, distribute software, and publish research/educational resources.

1. What is data science?

To a general human being data science is often defined as the intersection of three areas shown below Venn diagram math/statistics, computation and a particular domain (e.g. biology) (Conway, 2010; Yu, 2014; Blei and Smyth, 2017). Implicit in this definition is the focus on solving specific problems in contrast with the type of deep understanding that is typical in academic statistics. The focus on problem solving is important because it explains differing judgments to be found on how to value contributions to the field. As stated in Cleveland (2001)



(The data science Venn diagram)

How to read the Data Science Venn Diagram, The primary colours of data: hacking skills, math and stats knowledge, and substantive expertise.

ISSN (Print): 2278-0793
ISSN (Online):2321-3779

Science should be judged by the extent to which they enable the analyst to learn from data... Tools that are used by the data analyst are of direct benefit. Theories that serve as a basis for developing tools are of indirect benefit. We define data science as the union of six areas of greater data science which are borrowed from David Donoho's article titled "50 Years of Data Science" (Donoho, 2017).

1. Data gathering, preparation, and exploration
2. Data representation and transformation12
3. Computing with data
4. Data modeling
5. Data visualization and presentation (This includes both databases and mathematical representations of data.)
6. Science about data science (For example, reproducible research would fall under this category and point 5.)

The purpose of defining data science in this way is to (A) better capture where people who work with data spend their time/effort and (B) put more focus on the value of each tool for providing insights. This definition is given in contrast to the current, perceived state of statistics. A number of other statisticians have proposed similar definitions using different terminologies and ontology. The term data science, both the literal string and the broader idea that it conveys, originates from statisticians. In a 1993 essay titled "Greater or Lesser Statistics: a Choice for Future Research" statistician John Chambers wrote (Chambers, 1993).

Greater statistics can be defined simply as everything related to learning from data, from the first planning or collection to the last presentation or report. Lesser statistics is the body of specifically statistical methodology that has evolved within the profession generally statistics as defined by texts, journals, and doctoral dissertations. Greater statistics tend to be inclusive with respect to methodology, closely associated with other disciplines, and practiced by many outside of academia and often outside professional statistics. Lesser statistics tends to be exclusive, oriented to mathematical techniques, less frequently collaborative with other disciplines, and primarily practiced by members of university departments of statistics etc.

2. CRITIQUES OF STATISTICS:

Firstly we summarize the critiques of statistics as: too much theory, not enough computation. We believe theory is important; however, too much theory at the expense of other things is a detriment to the field. Statistics was primarily developed to help people deal with pre computer data problems like testing the impact of fertilizer in agriculture, or figuring out the accuracy of an estimate from a small sample.

Data science emphasizes the data problems of the 21st Century, like accessing information from large databases, writing code to manipulate data, and visualizing data. Another popular, but incorrect belief is that statistics is not concerned with big data.

As points out, statisticians have in fact always been interested in large data computation. For example, the word statistics came about from work on census data, which have been around for centuries and are large even by today's standards. The principle of sufficiency is of course a mechanism to deal with large data sets efficiently. The point here is that these pursuits are a part of statistics, but are perhaps considered specialized as opposed to mainstream (e.g. in terms of publications in flagship journals, undergrad/graduate education, etc). Statistics, on


ISSN (Print): 2278-0793
ISSN (Online):2321-3779

the other hand, has not changed significantly in response to new technology. The field continues to emphasize theory, and introductory statistics courses focus more on hypothesis testing than statistical computing... For the most part, statisticians chose not take on the data problems of the computer age.

The second is the statement that "statistics courses focus more on hypothesis testing". This makes the statistical outsider's mistake of thinking that statistics is a set of recipes for doing data analysis. It misses the deeper truth understood by people who practice data analysis: when properly taught, statistics courses teach an important way of thinking called the scientific method. The main idea is that to be really sure of making actual discoveries (as opposed to finding spurious and ungeneralizable sampling artifacts) scientists should first formulate a hypothesis, then collect data and finally analyze. One can be forgiven, however, for mistaking statistics as a set of recipes. Too many people interact with statistics exclusively via a standard Statistics type class which may in fact treat statistics as a handful of formulas to memorize and steps to follow. While we believe the material taught in these courses is vital to doing science, it is perhaps time to rethink such introductory classes and teach data before teaching statistics.

3. SOME PRINCIPAL COMPONENTS OF DATA SCIENCE:

(i)Prediction vs. Inference:

Prediction vs. inference is a spectrum. Many complicated problems have well defined sub problems which are closer to one end or the other end. The distinction we are trying to make here is maybe better described as engineering vs. science. Engineering is the business of creating a thing that does something. Science is the business of understanding how something works.

(ii)Empirically vs. theoretically driven:

Most quantitative fields of study do both theoretical and empirical work for example. theoretical vs. experimental physics. Within statistics, we might contrast exploratory data analysis vs. confirmatory analysis i.e. searching for hypotheses vs. attempting to confirm a hypothesis.

(iii) **Problem first vs. hammer looking for a nail:**

Researchers take a hammer looking for a nail approach; the researcher has developed/studied a statistical procedure and then looks for problems where it might be applicable. Other researchers aim to solve some particular problem from a domain. Note that the former approach is strongly correlated with, but not equivalent to theoretical research.

(iv) **The 80/20 rule:**

The basic idea of **80/20 rule** is that the first reasonable thing you can do to a set of data often is 80% of the way to the optimal solution. Everything after that is working on getting the last 20%.These includes data visualization, exploratory data analysis, data mining, programming, data storage/processing, computation with large datasets and communication.

As data analysis becomes more valuable, existing statistical theory and methodology also become more valuable. Criticisms of statistical theory are largely about ignoring other, less mathematically glamorous areas of statistics.

ISSN (Print): 2278-0793
ISSN (Online):2321-3779

4. EDUCATION:

A number of people have written about updating the statistics curriculum in ways which better reflect the broader definition of data science and the skills required for doing data analysis. A number of programs which have embraced these recommendations have proven to be successful such as the Johns Hopkins Data Science Specialization on Coursera. (Kross et al, 2017) .We observes three takeaways from this literature that is more computation, more data analysis and the use of open source material. Communication is another area worth highlighting in this education section because of its importance and ubiquity. The modern data analyst is expected to communicate across a number of different media, written papers/ reports, static/dynamic visualizations, online via creating a website/blog, through code, etc.

5. MORE COMPUTATION:

Computation comes into data analysis in a number of ways, from processing data to fitting statistical models to communicating the results. With the large number of technologies involved with data analysis (programming languages, visualization software, algorithms, etc) one often feels a bit overwhelmed at what one might be expected to know. It is infeasible to know everything. This is where updating the statistics education curriculum is critical. There is probably some rough core set of computational knowledge every statistician should have. Once the fixed cost of learning the core computational curriculum is paid, the marginal cost of learning additional computational skills will go down.

6. PEDAGOGY:

As many of the above references discuss, the current statistics curriculum often lacks data analysis (Tukey, 1962; Nolan and Temple Lang, 2010). Real data analysis makes the discipline more concrete to students. Focus on solving a real problem can be engaging to students who might otherwise find the subject boring. Teaching data analysis is challenging but it's challenging in the way that teaching the practice of engineering or using the scientific method is challenging. By not giving students practice doing data analysis for a real problem, the statistics curriculum may encourage students to view statistical methodology as a hammer to be procedurally applied to data. It's well established in engineering and the physical sciences that students should get some practical experience doing the thing during their education: why does the same principle not apply more often statistics at the undergraduate and graduate level? When teaching statistical modeling it might be more effective to first introduce the model (e.g. linear/logistic regression) in terms of a predictive context instead of the traditional inferential context.

Finally, we suggest teaching data before statistics in introductory statistics courses. In other words, we should teach exploratory analysis before inferential analysis. This would involve teaching programming, data visualization and manipulation before teaching hypothesis testing.

ISSN (Print): 2278-0793
ISSN (Online):2321-3779

CONCLUSION:

So far the arguments in the above paper have been about providing value to society by broadening the discipline in technical ways. Equally as important is increasing diversity in statistics by encouraging women and underrepresented minorities to join and stay in the discipline.

We return to the question of whether data science and statistics are really two different disciplines. If statistics is defined as the narrow discipline described by the quote from John Chambers then the answer is yes. However, if statistics embraces the broader idea of greater data science by putting more focus on computation in education, research and communication then we argue the answer is no.

REFERENCES:

[1]Alivisatos P (2017) Stem and computer science education: Preparing the 21st century workforce. Research and Technology Subcommittee House Committee on Science, Space, and Technology

[2] Anderson C (2008) the end of theory: The data deluge makes the scientific method obsolete. Wired magazine 16(7):16–07

[3] Aravkin A, Davis D (2016) A smart stochastic algorithm for nonconvex optimization with applications to robust machine learning. arXiv preprint arXiv:161001101

[4] Association AS, et al (2014) Curriculum guidelines for undergraduate programs in statistical science. Retrieved March 3, 2009, from http://www amstat org/education/curriculumguidelines cfm

[5] Barnes N (2010) Publish your computer code: it is good enough. Nature News 467(7317):753–753

[6] Barocas S, Boyd D, Friedler S,Wallach H (2017) Social and technical trade-offs in data science

[7] Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence 35(8):1798–1828

[8] Bhardwaj A (2017) what is the difference between data science and statistics? URL https://priceonomics.com/whats-the-difference-between-data-science-and/

[9] Blei DM, Smyth P (2017) Science and data science. Proceedings of the National Academy of Sciences 114(33):8689–8692

[10] Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Advances in Neural Information Processing Systems, pp 4349– 4357

[11] Bottou L, Curtis FE, Nocedal J (2016) Optimization methods for large-scale machine learning. arXiv preprint arXiv:160604838

[12] Breiman L, et al (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science 16(3):199–231

[13] Buckheit JB, Donoho DL (1995) Wavelab and reproducible research. In: Wavelets and statistics, Springer, pp 55–81

[14] B¨uhlmann P, van de Geer S (2018) Statistics for big data: A perspective. Statistics & Probability Letters

[15] B¨uhlmann P, Meinshausen N (2016) Magging: maximin aggregation for inhomogeneous large-scale data. Proceedings of the IEEE 104(1):126–135

ISSN (Print): 2278-0793
ISSN (Online):2321-3779

[16] B¨uhlmann P, Stuart AM (2016) Mathematics, statistics and data science. EMS Newsletter 100:28–30

[17] Members RP (2017) the r project for statistical computing. URL https://www.r-project.org/

[18] Hardin J, Hoerl R, Horton NJ, Nolan D, Baumer B, Hall-Holt O, Murrell P, Peng R, Roback P, Temple Lang D, et al (2015) Data science in statistics curricula: Preparing students to "think with data". The American Statistician 69(4):343–353

[19] Hicks SC, Irizarry RA (2017) A guide to teaching data science. The American Statistician (just accepted):00 00

[20] Hooker G, Hooker C (2017) Machine learning and the future of realism. arXiv preprint arXiv:170404688

[21] Huber PJ (2011) Robust statistics. In: International Encyclopedia of Statistical Science, Springer, pp 1248–1251