Received: November 12, 2017
Revision received: March 21, 2018
Accepted: March 26, 2018

www.estp.com.tr
DOI 10.12738/estp.2018.5.023 ♦ October 2018 ♦ 18(5) ♦ 1236-1245

Research Article

Impacts on Statistics Education in Big Data Era

Yanling Li[1] Chunyan Huang[2] Ling Zhou[3]

North China University of Water Resources and Electric Power *North China University of Water Resources and Electric Power* *North China University of Water Resources and Electric Power*

Abstract

In recent years, the development of computer and Internet has brought the amount of information to an unprecedented extent, and the world has entered the era of big data. The emergence of big data has epoch-making significance for statistics. Featuring diversity, scale and high speed, the big data makes up for the disadvantages (high cost and high error) of statistics. However, this does not mean that the age of statistics is over, and the processing of big data still needs to rely on statistical methods. The age of big data brings both opportunities and new challenges to the development of statistics. This paper analyzes the relationship between big data and statistics, discusses the change of statistics teaching mode and teaching concept in the era of big data, analyzes the impact of big data on statistics and summarizes some problems existing in the research and teaching about big data.

Keywords

Big Data • Statistics Education • Teaching Mode • Challenges and Opportunities

[1]**Correspondence to:** Yanling Li (PhD), Department of Mathematics & Information Science, North China University of Water Resources and Electric Power, Zhengzhou 450045, China. Email: liyanling@ncwu.edu.cn
[2]Department of Mathematics & Information Science, North China University of Water Resources and Electric Power, Zhengzhou 450045, China. Email: huangchunyan@ncwu.edu.cn
[3]Department of Mathematics & Information Science, North China University of Water Resources and Electric Power, Zhengzhou 450045, China. Email: zhouling@ncwu.edu.cn

Citation: Li, Y. L., Huang, C. Y. & Zhou, L., Impacts on Statistics Education in Big Data Era. *Educational Sciences: Theory & Practice*, *18*(5), 1236-1245. http://dx.doi.org/10.12738/estp.2018.5.023

Along with the popularization of Internet e-commerce, the wide use of smart mobile devices, increasingly frequent social networking activities, and the rise of cloud computing and Internet of things technology, data starts to grow geometrically, and the world steps into the era of big data. According to IDC's Data Universe, the global data volume is 0.5ZB in 2008 (1ZB equals 1 trillion GB and 1.8ZB is equivalent to 1.8 billion 1TB mobile hard drives). The data volume is 1.2ZB in 2010 and humans entered the ZB era officially. What is more striking is that the global data volume will continue to grow at a rate of more than 40% annually until 2020. It doubles almost every two years and is expected to break 35ZB by 2020 (Ridgway, 2016; Hardin, Hoerl, Horton, & Nolan, 2015). Therefore, how to apply and analyze the increasing data information has become a problem that both domestic and foreign scholars are keen to think about and explore (Lynch, 2008; Zwick, 2015). In this context, the teaching of statistics shall also adapt to the new changes of the times (Rifkin, 2012). In the course of teaching, teachers need to combine the current big data era with their original professional knowledge, constantly explore the teaching content, and continuously improve the teaching method.

Scholars Zhu Jianping et al. (Zhu & Zhang, 2016) defined and discriminated the concept of big data, analyzed the change of modern information technology and Internet technology, and emphasized that higher education must pay attention to big data utilization, and that it is necessary to change the learning mode in the field of higher education, promote personalized education to assist students, and promote the communication of scientific research, which provides a new thinking for the development of higher education in the era of big data. Lin Hong et al. (Lin and Li, 2014) suggested that statistics must be integrated with computer science to embrace the arrival of big data era.

The core of big data is data, and data is the object of statistical study. The key to find valuable information from big data is to make accurate statistical analysis of the data (Bughin, Chui & Manyika, 2010). The true value of data is like the icebergs floating in the ocean. At first sight, only the tip of the iceberg can be seen, while the vast majority is hidden beneath the surface. In the era of big data, a lot of data is in the state of "dormancy", and to unlock and release the hidden value of these data, unremitting efforts need to be made by statisticians with the aid of a new generation of methods and tools. Because of the emergence of big data, the definition, thinking mode and function of statistics are different from those of traditional statistics. Undoubtedly, with the advent of the big data era, statistics has entered a new stage of development (Geng, 2014).

Statistics and big date

Statistics, as a discipline, has a history of more than 300 years. According to the evolution sequence of statistical methods and history, the history of statistics can be divided into three stages, namely, classical statistics period, modern statistics period and contemporary statistics period. The germination of classical statistics can be traced back to the middle of the 17th century. Modern statistics developed from the late 18th century to the late 19th century (Statistical…, 2017). The development of science and technology in the 20th century is much faster than before, modern statistics with descriptive method as the core can no longer meet the demand. The focus of statistics turns to inferential statistics and the period of contemporary statistics has come (Sprent, 2003). During this period, statistics has been widely applied in toxicology, molecular biology, clinical

tests and other biomedical fields, and the development of these fields contributes to the continuous innovation of statistical methods. Non-parametric estimation, MME algorithm and other methods were born at the right moment (Wagstaff, 1991). As one of the fairly abstract concepts, people have different opinions about the concept of big data. No matter how big data is defined, the key to find valuable information from large data is to make correct statistical analysis of the data (Maurer, 2015). In the era of big data, data can be generated whenever and wherever possible. Not only the data collection is dynamic, but the data storage technology and data processing technology are also updated at any time, which means that the statistical tools processing the data will also change.

During the 40-day Spring Festival travel period in 2017, the total number of trips reached 2.7 billion. In the short-term migration which has the largest scale in human history, where did the crowd come from and where did they go？Which lines are the most popular？It can be seen from Baidu migration website that every jumping point on the map of China is a place where people start and arrive (Figure 1). In the past, these questions may be difficult to answer precisely. However, the "big data" technology enables people to see the whole picture of Spring Festival travel season from the labyrinth of data.

Figure 1. China's Spring Transportation Map in 2017

Driven by the real demand, statistics develops gradually along with the change of demand and data. In the era of big data centering on data information, the development of various fields needs to derive power from big data, which undoubtedly generates a lot of demand for statistical analysis. According to the discipline characteristic and its historical evolution, it is not hard to see that, in the era of big data, based on the characteristics of big data, statistics shall take serving and meeting the demand of various fields as the goal, continuously innovates and develops data analysis method and theory.

Impacts of big data on statistics teaching

Changes of teaching mode and teaching concept

In history, human teaching has undergone two major changes. The first change is marked by the establishment of private school by Confucious and lectures given by Socrate, which are both pioneering in the history of education. The second change occurred in the 16th century when Comenius, the Czech educator

introduced industrialization revolution into teaching and created a classroom learning system. Teachers carried out indoctrinating teaching with a one-to-many mode in a fixed place. This teaching mode is still in use today. In the 21st century, information has become the education mode in the information age. The information technology develops rapidly. The emergence of MOOC has become a hot topic in domestic and foreign teaching fields. It really triggers a learning and education revolution, which is changing the traditional classroom teaching mode. Statistics has rich material resources. In the face of the new wave of education revolution led by MOOC, and with the development of big data, how the teaching model will change deserve our attention and consideration (Naimi and Westreich, 2013). In addition, with the aid of Udacity, Coursera, edX and other online education platforms, many high-quality courses offered in the world first-class universities will be open free online, so that high-quality education resources are provided at a low cost to anyone in the world who is willing to accept and learn. This kind of education and teaching development ideas based on big data will have great and far-reaching impact on traditional education. In the age of big data, the knowledge acquisition will no longer be confined to the classroom. Compared with the step-by-step teaching mode in traditional classroom which confines time and space, online teaching enables students to learn more autonomously, because they can arrange their study time and place freely. The online education platform connects learners with excellent education resources to enable anyone to learn without obstacles. From the perspective of the teaching form, it belongs to remote education; from the perspective of teaching method, it is a kind of open course teaching which utilizes the computer Internet information exchange and the big data information mining function. This kind of teaching mode is convenient for students to review and understand the content learnt previously. In addition, there are some interactive micro videos, which can effectively reduce the fatigue caused by online learning, and help students focus attention and improve their learning efficiency.

Online education provides students with high-quality education resources of world first-class universities for free, so university teaching will face more intense competition and challenges. How will the role of university teachers change when students can access the first-class teaching resources on the Internet for free? Some of the excellent teachers may get better development through providing resources of good quality, while some teachers provide students with tutoring and help beyond the online education. Some teachers will deflect themselves from the basic education and devote to the work of scientific research. Therefore, the research function of universities will be strengthened, which will better promote the rapid development of scientific research.

Impacts on statistical subject

The research object of statistical discipline is the quantitative features and quantitative relation of objective things. Traditional statistics believe that data is mainly derived from the numerical value of tests, experiments or surveys. In the era of big data, not only the quantity which is measured with structural data can be taken as the statistical research object, but the semi-structured and unstructured data that cannot be measured by quantitative relation can be taken as the research object. Big data expands the object of statistical study.

The concept of sample in statistics in the big data era. As is known to all, traditional statistics cannot be separated from samples. Samples are a part of individuals which are actually observed or investigated in the study. An available sample must be able to reflect the overall situation correctly. In the era of big data, the

concept of samples is no longer so simple, and data is divided into two types: static data and dynamic data (Fan, Han & Liu, 2014). For the static data, samples equal to the population. Therefore, there is no need to extract samples and test the availability of samples. The cost is lowered and the error is reduced since the population itself can more accurately reflect the population. The dynamic data changes over time. At this time, the population is the sum of all the data in the history. Our analysis object is the "sample". Different from the concept of traditional sample, the "sample" here is not limited to the randomly chosen data and can be the selected data related to the analysis purpose.

The sampling thought deepens in statistics in the big data era. Schoenberg et al. mentioned in the Age of Big Data (Saha and Srivastava, 2014) that "In the big data era, the population is required and sampling is not needed". In the big data era, all the data that can be recorded with modern information technology can be provided. In addition to the common information and universal laws between things, personalized feature information can also be provided. There may be systematic error in big data generation and acquisition under the artificial design framework (Gal and Ograjenšek, 2017), therefore the sampling method of big data needs to be studied. Not all the "data" is required to serve as "sample". No matter how big the pot is, as long as it is stirred evenly, the flavor can be known through tasting a small spoon of the content. For the large data stream environment, it is necessary to extract a sample that can satisfy the statistical purpose and precision from the steady stream of data and to study an adaptive, sequential and dynamic sampling method.

Change of the analytical thinking

Data in the traditional sense is structured data, which contains specially designed qualitative data and the quantitative data that can be presented with conventional statistical index or diagram. They have regular structures and standards (Yang, Chen & Zheng, 2017). Featured by diversity, the big data includes not only structured data, but also unstructured data, semi-structured data, or heterogeneous data, that is, all the signals that can be recorded and stored. The big data cannot necessarily be fully expressed with traditional statistical indicators, so the data identification and analysis methods are more diversified.

In the long-term development, statistics has formed a unique way of thinking, and the traditional statistical analysis process is divided into three steps, namely, determination on the nature, quantification and determination on the nature again. Firstly, the statistical direction is found through experience-based judgment, then the data is quantified, analyzed and processed, and finally, the conclusions are drawn according to the results (Chen and Xin, 2017). In the big data era, the statistical analysis process is "quantification-determination on the nature". The basic work is to find the "quantitative response", and find valuable data directly from various "quantitative responses". Next, the data features and quantitative relations are found out through analysis, and then judgment and decisions are made accordingly.

In addition, the train of thought of the traditional statistical empirical analysis is "hypothesis-verification". In the era of big data, the idea of the empirical analysis is "discovery-summary". In order to understand the research object in a more comprehensive and deeper way, the data needs to be integrated to look for relations, and explore rules, and then summaries are made and conclusions are drawn. This way is conductive to discover

more unexpected "discoveries". The traditional statistical inference analysis process is based on the distribution theory, and the inference is made to the population on the premise of guaranteeing probability (Strang, 2016). The processing of big data becomes the judgement of probability based on the actual distribution and the features of the population. It is not required to infer the general characteristics according to the distribution theory, but to draw inferences based on the calculation method.

In the era of big data, the requirements of enterprises and government for statisticians are constantly improving. Therefore, the teaching of statistics should also keep pace with the times and seek development in reform.

Development of statistics brought by big data

According to Professor Gary King of Harvard University, big data is a revolution, a huge data resource enables various fields to start the process of quantization, regardless of academic world, business or government, all areas will begin this process." Today, we have entered the information society. In the face of the era of big data, cloud computing, Internet of things, mobile terminals and wearable devices are highly developed and integrated. Regardless of who you are, and whether you like it or not, you need to deal with the data. You are either generating data or accepting data. Big data is bound to bring new development to statistics (Gillborn, Warmington & Demac, 2017).

The quality of the statistics has been improved.

The advantage of statistics lies in that it can see big things through small ones. For big data, more and more data even the conceptual data can be utilized. The data limitation factors have become a history. Statistics can collaborate with big data, not only see big things through small ones, but also making complicated things simple. On the basis of big data, the applicability, timeliness and accuracy of statistics are greatly improved. In the era of big data, the collected statistical information is more in line with the needs of users. The time needed for statistical investigations will be shortened. Because of the comprehensiveness of large data samples, the accuracy of the statistical results can be guaranteed through reducing the human error in the statistical process.

The cost of statistics has been reduced

Seen from the data collection method, traditional statistical data collection methods mainly rely on surveys, such as questionnaires, telephone interviews, or inquiring statistical reports. The accuracy of these methods cannot be guaranteed and the statistical cost is rather high (Geng, 2014). In the era of big data, data is obtained through information network, mobile communication, etc. Therefore, from the perspective of various elements of the statistical cost, the statistical cost will fall dramatically in the big data era and larger scale and more accurate data will be obtained.

The statistics system extends

In the era of big data, the development of statistics should be viewed with development and dialectic vision, and statistics should construct a new discipline system under the framework of big data. It is necessary to integrate the thought and methods of population statistics of the big data into the statistical discipline system. In the teaching content of statistics, the sample statistics and the population statistics are combined together. The sample statistics requires the sample to correctly represent the population, so the population observation unit must be homogeneous. This ideal situation is not easily achieved in real life, yet the population statistics based on big data can compensate for this deficiency.

The function of statistics expands

The traditional statistics are mainly applied in industrial and department statistics due to the impact of cost, concept and other factors, providing service to the industries and departments for making and improving policies. In the era of big data, statistics can not only get more rapid development in the field of statistics, but also make the statistical principle and method applied to other disciplines, such as finance, medicine, computer, etc., so that statistics can play a greater role.

The demand for statistics specialty students increases

Big data plays a significant role in improving the employment of statistics major students. Nowadays, countless industries, such as governments, businesses and individuals are all hoping to extract valuable gold from the gold mine of big data. However, only knowing the industry knowledge is not enough for data mining. The industry knowledge needs to be combined with the data analysis technique. Therefore, the statistical staff and data analyst develop. In the era of big data, they take advantage of their own professional advantages, changing various types of data into valuable information. They have been playing a positive role in promoting the development of all walks of life, and can improve the industry experts' level of thinking and working efficiency. In the future, the role of statisticians and data analysts will not be underestimated, and their position will surely be greatly improved.

Problems existing in statistical education

Big data brings new development opportunities to statistics, and also brings some challenges to traditional statistics. Therefore, what is the future of statistical science in the age of big data? What big data problems need to be solved by statistics? These are the concerns of the statisticians and the teaching staff.

The development of classical statistics in the age of big data

Many traditional classical statistics methods, ranging from theory to practice, have performed well after the long-term test in different fields. However, the direct application of these methods in the age of big data will cause certain problems. It is not appropriate, or even a loss, to discard these classical statistical methods directly.

How to improve these traditional classical statistics methods combining with the high-speed calculation method and corresponding software and hardware environment to apply them to big data is a question worth pondering.

Confluence analysis of the multi-source heterogeneous big data

Big data is multi-source heterogeneous data covering different ranges. In the age of big data, data from multiple sources often appear in the description of the same object or problem. In order to integrate various data, the data source, data acquisition mode and data description should be formalized to support the data analysis. It is more convenient to collect data in the big data era. Through the effective integration of data, on the one hand, more abundant information can be obtained, and on the other hand, data from different sources confirm with each other, which can verify the authenticity and accuracy of the information. In the big data environment, many datasets no longer have keywords that mark individuals. Traditional relational database connection methods no longer apply, so it is necessary to explore the method to take advantage of the overlapping projects between databases to combine different databases and to integrate the data of multiple different variable sets into a large database containing complete variable set based on the conditional independence between variables. Of course, the confluence analysis and modeling of multi-source heterogeneous big data is also one of the important development directions of statistics.

The marginal effect problem of big data

The big data era provides people with an open information system to collect data from various information collection devices. In the big data environment, people who collect data may not know what the data user will do with the data in the future, the data modeling person perhaps does not know how the data is obtained and the model user may not know what kind of data the model is from. Therefore, people inevitably interpret models overly according to their own intentions, going beyond the information range contained in the original data. In other words, it is not that the more the better for the data. The information value generated by big data has marginal diminishing effect, i.e. when the data volume is large enough to a certain extent; the information provided no longer increases, while the data collection and handling cost is rising constantly. Meanwhile, the noise contained in the big data affects the information extraction. Therefore, in the era of big data, the amount of data should not be blindly pursued, and the balance between cost and utility should be considered to choose an optimal data volume. Therefore, it is necessary to discuss the marginal effect problem of big data in modeling from statistical perspective.

Conclusion

The advent of the era of big data will certainly lead to the significant changes in the teaching of higher education. The high-dimensional mass data is also featured by diversity and high speed. The practical application and data drive the development of statistics. The teaching staff has a responsibility to provide every student proper teaching program, so that teachers can obtain the teaching feedback more effectively and objectively, improve the teaching quality and cultivate more big data talents, thus promoting the development

of statistics in the age of big data. Although a lot of researches and discussions have been carried out to the data analysis concept and thinking reform, people's understanding of big data always takes some time.

The emergence of big data has epoch-making significance for statistics. Featuring diversity, scale, large quantity and high speed, the big data makes up for the disadvantages (high cost and high error) of statistics. However, this does not mean that the age of statistics is over. The search, clustering and classification of big data still require statistical methods; therefore, big data cannot be separated from statistics. The advent of the era of big data improves the quality of statistics, lowers the statistical cost, makes statistics play a bigger role in a wider range, extends the statistical disciplines and improves the status of statistics in natural science and social science, which are the opportunities brought by the big data to traditional statistics. At the same time, the traditional statistics are also facing challenges. The statistics teachers are required to change their understanding of some basic concepts, establish a series of new data analysis methods, and effectively utilize the big data resources. We shall firmly seize the opportunities brought by big data, actively respond to the challenges and organically integrate the big data with statistics to keep the exuberant vitality of statistics in its future scientific development.

References

Bughin, J., Chui, M., & Manyika, J. (2010). Clouds, big data and smart assets: Ten tech-enabled business trends to watch. *McKinsey Quarterly*, *56*(1) ,75-86.

Chen, W., & Xin L. (2017). Art S O, et al. The application of big data in college moral education. *State Academy of Forestry Administration Journal*.

Fan, J., Han, F., & Liu H. (2014). Challenges of big data analysis. *National Science Review*, *1*(2), 293- 314. https://dx.doi.org/10.1093/nsr/nwt032

Gal, I., & Ograjenšek, I. (2017). Official statistics and statistics education: Bridging the gap. *Journal of Official Statistics*, *33*(1), 79-100.

Geng, Z. (2014). Opportunities and challenges in the age of big data for statistics. *Statistical Research*, *31*(1), 5-9.

Gillborn, D., Warmington, P., & Demac, K. S. (2017). QuantCrit: education, policy, 'Big Data' and principles for a critical race theory of statistics. *Race Ethnicity & Education*, *21*(2). 158-179. https://dx.doi.org/10.1080/13613324.2017.1377417

Hardin, J., Hoerl, R., Horton, N. J., & Nolan D. (2015). Data Science in Statistics Curricula: Preparing Students to "Think with Data". *American Statistician*, *69*(4), 343-353.

Lin, H., & Li, Z. J., (2014). The integation and development of economic statistics computer science and information science in the era of big data. *Economic Statistics*, *3*(2), 10-17.

Lynch, C. (2008). Big data: How do your data grow, *Nature*, *455*(7209), 28.

Maurer, T. K. (2015) Applications of technology and large data in statistics education and statistical graphics. *Dissertations & Theses - Gradworks*, https://dx.doi.org/10.31274/etd-180810-4191

Naimi, A. I., & Westreich, D. J. (2013). Big Data: A revolution that will transform how we live, work, and think. *Mathematics & Computer Education*, *47*(17), 181-183.

Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical*, *84*(3), 528-549. https://dx.doi.org/10.1111/insr.12110

Rifkin, J. (2012). The third industrial revolution: How lateral power is transforming energy, the Economy,and the World. *New York: Palgrave Macmillan*. *Survival*, *2*(2), 67-68.

Saha, B., & Srivastava, D. (2014). Data quality: The other face of Big Data, *IEEE, the 30th International Conference on Data Engineering*, *New York: IEEE*, 19-46.

Sprent, P. (2003). Modern medical statistics: A practical guide. *Journal of the Royal Statistical Society*, *167*(1), 183-198. https://dx.doi.org/10.1111/j.1467-985X.2004.298_8.x

Statistical Methods in Big Data Working Group, (2017). Reflections on the Positioning of Statistics in the Big Data Era. *Statistical Research*, *34*(1), 5-11.

Strang, K. D. (2016) Beyond engagement analytics: which online mixed-data factors predict student learning outcomes. *Education & Information Technologies*, *22*(3), 1-21. https://dx.doi.org/10.1007/s10639-016-9464-2

Wagstaff, A. (1991). Wagstaff QALYs and the equity-efficiency trade-off. *Journal of Health Economics*, *10*(1), 21-41. https://dx.doi.org/10.1016/0167-6296(91)90015-F

Yang, J., Chen, T., & Zheng, L. S. (2017). The Influence of the big data age on the teaching of statistics, form Education Teaching Forum.

Zhu, J. P., & Zhang, Y. H. (2016). Reflections on conventional statistics in the big data era. *Statistical Research*, *33*(2), 3-9.

Zwick, M. (2015), Big data in official statistics. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*, *58*(8), 838-843. https://dx.doi.org/10.4018/978-1-5225-2512-7.ch011