Statistics in the Age of Big Data: Opportunities and Challenges

Guoli Liu
Department of Statistics, University of Wisconsin-Madison, Madison WI 53703, US
liuguoli1703@163.com

Abstract. The era of large data will change the way we understand and form statistical information. To store, integrate, process and analyze massive amounts of data can be viewed as a new data analysis method. This kind of intrinsic nature based on data relations determines the inevitable relationship between large data and statistics. Big data provides both opportunities and challenges for the development of statistics. The opportunities are reflected in: the improvement of statistical quality, the decline in statistical costs, and the expansion of the field of statistical action. The challenges are reflected in: the standard of large amounts of samples to be adjusted, the sample selection criteria and form to be re-determined, the statistical software to be upgraded and developed.

Keywords: Big data; statistics; opportunities.

1. Introduction

For the definition of large data, the practitioners from different disciplines, different industry will certainly have different understandings. Some scholars use 4V (Volume, Variety, Velocity and Value) to describe the characteristics of large data [1]. Large data in the data size, data complexity and the speed of production are greatly beyond the traditional statistical data form. From the view of disciplinary, the approaches large data provides to the massive data storage, integration, processing and analysis can be regarded as a new data analysis method.

Traditional data analysis methods are usually statistical research areas. Through the collection, collation and analysis of statistical data, statistics achieve the purpose of exploring the inherent nature of the data. Traditional statistical theories and methods are almost based on the samples. This is because in the small data age, by the constraints of tools and capacity for collecting and processing data, it is almost impossible to collect all the data related to the problem. So the traditional statistical theory and methods are usually based on the performance of random samples to represent the performance of the whole. However, even if the optimal sampling and statistical analysis method is selected, the samples can only restore the characteristics of the whole in a certain aspect or a few aspects, and cannot restore them in all aspects.

With the development of technologies such as global positioning systems, sensors, and the Internet, much data that was previously difficult to collect can now be easily and efficiently collected on a large scale. Cloud computing makes large-scale data processing possible [2]. The analysis based on massive data allows us to obtain a new vision that cannot be achieved without the use of samples, which provides an unprecedented opportunity for statistical development. At the same time, how we combine the characteristics and needs of large data with traditional statistical methods to improve and find new statistical theories and methods to better adapt to the development requirements of large data age, which is the challenge the large data time must meet.

2. The Difference between Big Data Statistics and Traditional Statistics

2.1 Differences between Sample Statistics and Full Sample Statistics.

Statistics depend on sample statistics. A sample is taken from the population according to a certain probability and aggregated as a whole. And random sampling has costs, such as the time cost, the capital costs, and social relations and so on. In the case where the size of the sample is limited, the larger the overall number is, the larger the estimated error is. That is an unavoidable defect in the sample statistics. In the large data age, storing and processing data are no longer as difficult as in the past, almost all of the transaction information will be digitized and stored on the computer, which

makes the whole data possible. Large data analysis is analyzing all the data, that is, all the relevant data will be analyzed, so that the sample is equal to the overall statistical [3]. Full data can not only bring us higher accuracy, but also let us see some details which cannot be found before. Large data analysis allows statistical analysis to see more clearly the details of the sample that cannot be revealed.

2.2 No Longer Have to Find a Causal Relationship.

In the traditional statistical work, finding a causal relationship is a habit for a long time. Even if it is difficult to determine causality, we are still habitual to find the reason. However, some scholars believe that, strictly speaking, statistics cannot test the logical causal relationship. For example, according to the statistical results, it can be said that the incidence of lung cancer in smokers is several times higher than that of non-smokers. But the statistical results cannot draw the logical conclusion of smoking carcinogenesis.

In the era of large data, we no longer need to focus on the causal relationship between things, and more is looking for the correlation relationship between things. Although the traditional statistical work also studied the correlation relationship, the nature of the correlation analysis between large data is that the starting point is not "hypothetical causal relationship" and it is from a large database which exists in reality. By analyzing the correlation relationship between the data, artificial assumptions can be excluded to explore the deep meaning of data.

3. Big Data Provides Statistical Opportunities

The existence of massive data allows us to use more data when we deal with problems with the statistical methods. And even in some cases the whole data can be used. The data is no longer statistically significant, and the statistical accuracy and the accuracy of fitting prediction based on large data can be greatly improved, and many details that cannot be found on the sample statistics can be found.

3.1 Improve the Quality of Statistics.

Reasonable use of large data is conducive to the improvement of statistical quality, mainly in three aspects: timeliness enhancement, error reduction and credibility enhancement. Traditional statistics are often hysteresis and exhibit low frequencies, and the timeliness of large data can compensate for the shortcomings of traditional statistical data, so that the timeliness of statistical data is enhanced. CPI to the consumer price index (CPI) statistics, for example, CPI released for the frequency, but generally there is a lag, such as China's CPI is usually in the month of 9 to release the last month's CPI. The "online price index" to market prices in real-time tracking and aggregation, to provide timely statistical information, and online price index can be raised from the month to the day even higher, to analyze the law of inflation.

At the same time, the extensive coverage of large data can greatly reduce the error of statistical results. We still take CPI as an example. The traditional price statistics includes a basket of goods, usually contains thousands of goods, and involves tens of thousands of survey sales outlets. And the kind and structure of the commodity should be adjusted with the social and economic development and the consumption structure of the people. The sample error and the human error are all big. And the "online price index" based on large data makes sampling no longer important. Statistical objects can be tens of thousands of goods, all online vendors and most of the offline sales outlets, and even cover all the samples, which significantly reduce the statistical error.

3.2 Statistical Cost Reduction.

Whether the survey is a census or a sample survey, the traditional survey method is mainly telephone interview, questionnaire, statistical reports and other common methods. However, these methods all have their inherent defects. And if you want to obtain a larger amount of data, the statistical cost will greatly increase. The telephone interview process is easily interrupted by the unilaterally hanging up, the recovery rate and availability of questionnaires are usually not high, and the manpower and material resources spent in the process of reporting the statistical reports layer upon layer will increase as the number of layers reported. While in the era of big data, many data can be obtained

through the network, mobile communication, etc., so whether from time or from the actual consumption of financial and material resources, the statistical cost of big data will be greatly reduced compared with the traditional statistical survey method, and the data size will be larger.

More importantly, big data can be reused. The data collected is no longer limited to a particular purpose, it can serve a variety of different purposes. With the increase in the number of times the data was used, the potential value of the data is also increasing, while the cost of data collection is fixed instead of changing with the number of data utilization. So the average cost per time will be significantly reduced as the number of reutilization increases.

3.3 Expanding the Field where Statistics Can Play a Role.

With the development of human records and measurement methods, many things that cannot be quantified before can now be translated into data that can be analyzed, which means that the problems that were previously unable to be analyzed statistically can now be processed using statistical methods. In the era of big data, large amounts of data are extracted from some areas that were previously considered not to be digitized. For example, the web users' health status information and merchant preference information may be extracted from their search history, and the credit records and property status of the relevant information may be extracted from the social network of users. As long as data can be obtained, you can use statistics to carry out data analysis. So with the development of big data, the field where statistics can play a role will expand.

4. Big Data Presents Statistics with Challenge

That statistics in future will be turned into big data is an inevitable trend. There is still some incompatibility between the existing statistics and big data. In order to actively respond to this trend, it is necessary to make a corresponding adjustment of the existing statistical theory and method and even some aspects of the complete innovation.

4.1 Adjustment of Large Sample Standards.

Statistics rely on sample statistics, and mainly study the relationship between the number of objective things and quantitative characteristics. Massive amounts of real-time electronic data are produced in the age of big data, which sample size is even large enough to cover the whole population, so it contains more information. For example, the traditional economic statistics are generally refined to the industry level or product level, but the development of e-commerce and the popularity of bar codes make the record specific to each transaction. Online electronic trading information, electronic business records and the records of electronic departmental administration provide a large amount of data for the statistical survey, providing a possibility for the expansion of the statistical sample size.

Under the age of big data," the sample is overall" will be a new trend, and the standard of large sample also need to be improved accordingly. It is generally considered that the number of samples greater than 30 is a large sample, whereas the number of samples smaller than 30 is a small sample [4]. Traditional statistics treats 30 as the large sample standard. In the face of the existence of multi-source heterogeneous, high noise and other characteristics of the big data resources, the standard is too low to remove the impact of interference information, resulting in the statistical results can only explain the law of the change of the phenomenon to a limited extent. On this basis, traditional statistics should make full use of massive data provided by the big data to enhance the diversification of data sources, thereby expanding the size of samples and to update the large sample standards at the same time to replace the original sample quantity with a larger sample size so as to meet the requirements of data accuracy in the age of big data.

4.2 Redetermination of Sample Selection Criteria and Form.

Traditional statistics rely on structured data, such as numbers, symbols, etc., but unstructured data (including text, images, audio and video, etc.) and semi-structured data (such as HRML documents) also contain mass information and statistical law. At present, more than 85% of the data collected from big data is unstructured and semi-structured data. Traditional relational databases are not qualified for processing these unstructured and semi-structured data. However, big data can standardize the data by

building unstructured databases, turning the unstructured data into structured data, so as to play the potential role of these diversified data. If the traditional statistics can break through the constraints of structured data, reduce the sample selection criteria and establish unstructured databases to make the statistical data base diverse, the statistical application will be greatly expanded.

4.3 Statistical Software Remains to Be Upgraded and Developed.

Traditional statistical data processing and analysis are based on statistical models and statistical software. The statistical model constructs the quantitative relationship between the different variables. And statistical software is a powerful tool for processing and analyzing data, it must rely on the user to import a series of data they collected which related to the variables. Common statistical software includes SAS, SPSS, Stata, and Minitab. Big data relies on the techniques of non-relational data analysis based on data centers. If big data can be fully used in statistical software, the data collection process of the statistical analysis can be simplified or even removed.

It can be predicted that if statistical software commonly used in statistics can be modeled on big data processing software to increase data storage and transmission technology on the basis of data processing and analysis, statistical software itself can form an intrinsic data center to realize data sharing and thus promote the application of big data in statistical software. The application not only requires storage capacity and technical support, but also has strict requirements for the representation of the data. For the same variable of the massive data, the original data must have the necessary unified signs, otherwise the data center was established on the basis of which will be difficult to identify data.

5. Conclusion

Based on the difference between large data statistics and traditional statistics, this paper expounds the opportunities and challenges brought by big data to statistics. Under the age of big data, the statistical theory system and the statistical practice must be reformed and innovated. But we should be soberly aware that big data are complementary rather than substitutes for traditional statistics, and traditional statistics based on sample statistics and predictive analysis still play a leading role in social statistics and economic analysis.

References

[1]. Mayer-Schnberger, V. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think. Viktor Mayer-Schnberger and Kenneth Cukier*. Houghton Mifflin Harcourt.

[2]. Hayes, B. (2008). Cloud computing. *Communications of the Acm, 51(7)*, 9-11.

[3]. Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy & Technology, 25(4)*, 435-437.

[4]. Steel, R. G. D., & Torrie, J. H. (1960). *Principles and procedures of statistics.* McGraw-Hill.