**Interpretation of the Concept of Big Data from the Perspective of Statistics**

**Ying MEI**\(^{1,a}\) and **Gui-ping LI**\(^{1,b}\)

\(^{1}\)Chuxiong Normal University, Chuxiong City, Yunnan Province, China

\(^{a}\)my21@cxtc.edu.cn, \(^{b}\)lgp@cxtc.edu.cn

*Corresponding author

**Keywords:** The Concept of Big Data; Statistical Data; Diversity; Efficiency.

**Abstract.** This big data era is full of wisdom and warmth thanks to the development and the technological innovation. From the perspective of statistics, ‘big data’ means population under certain conditions, and it can be extended, excavated and it also embodies the relevance and diversity.

**To Build a Theoretical Framework of Statistics**

Statistics was started and developed as a discipline since seventeenth Century. And each period has its different school of thought, representative figures, typical theories, and the generation of famous statistical methods. With the increasing development and expansion of the statistical theory and practice, theoretical framework of statistics is elaborately build through discussion, debate and even struggle, each branch of the system is characterized by unique, continuity, accumulation, and combined with other disciplines as well.

The object of statistics is data, hence it is an important issue to understand the new concept of ‘big data’ from the perspective of statistics.

**The Concept of Big Data from the Perspective of Statistics**

The wheel of history in advance, our life has been changed turn the world upside down. Thanks to the development of information technology, our shopping, traveling, businesses, sports, entertainments, social behaviors are all maintained through texts, photos, videos, and digital records. Not as the saying goes ‘birds in the sky to fly and sky does not know where’re they fly, fish in the sea to sway and sea does not know where’re they go’, where I’ve been, ‘someone’ records it in full for me by Travel Apps such as ‘Umetrip’. Life is full of indelible data traces and unforgettable data memories with the schedule, and the route map. This is the wisdom and warmth of data era thanks to the development and the technological innovation. According to the statistical definition, these traces and memories carrying our information so called ‘data’, is different from traditional data, we now call it big data.

‘Big data’ is a frequently mentioned, and often appears hazy word because different person holds different points and purposes of viewing problems, methods of the research fields are also different. So it is very hard to define it commonly in each subject category. Definition can be generally accepted is almost impossible.

Something deeper, we can interpret the definition from two sides, the ‘big’ side and the ‘data’ side.

**Interpret ‘big data’ from the ‘big’ side**

The conception of big can be understood from the following three aspects.

First, ‘big’ means complete, big data means all the data under certain conditions, and the volume of the data must reach a certain size. For example, we use a so called campus card through modern information technology to record the information of all school teachers and students, including every consumption record in canteens, supermarkets, shops, bookstores, cafes, as well as book lending
trace in the library, the comprehensive performance and the GPA, the duration on line and bathing, the scholarships, punishments and loan repayments. All traces of teachers’ and students’ campus life are recorded into the campus cards, and these variety kinds of intricate data can be comparative analyzed through algorithm, and some metrical rules can be find through the technical processing. This all traces within the scope of view are so called ‘big data’.

Secondly, ‘big’ means can be extended, fresh data can be continually absorbed, storaged and expanded once it happened. Such as every increasing consumption record including students’ information, transaction information and business information, can be absorbed, recorded and stored down into data warehouse by the school campus card, different from the traditional method of collecting data. It is because of the development of information technology can do it all beyond the traditional storage, processing and analysis capabilities, the concept of big data really grow up. It is also due to the technology development, the ‘expansion’ makes the ‘big data’ a meaningful word.

Thirdly, ‘big’ means can be excavated. The data carrying information embodies value, though the information is often accompanied by noise coexist in the data. Obviously, big data often contains the perplexing and large value of the information which deserves mining. We use the campus card example again, through mining useful information from the field symbols we can know the number of merchants who provide food is more popular, we can identify the consumer behaviors of teachers and students who eating in the cafeteria, we can also clear what factors (time, season, consumer preferences, personal background, distance, kind, taste, environment, service quality and level etc.) affect and how to influence consumer behavior. And they’re very helpful to give feedback, improve product quality, and reduce the waste of resources and so on. On the basis of the ‘big’ volume, the mining rules are very stable, which is the value of the information.

Finally, ‘big’ also embodies the relevance and diversity of data itself. Big data covers and contains everything. The biggest difference between big data from traditional data is online, the recording, calling, storage, calculation and analysis are almost done online which leads to the data is no longer confined to the typical structural form of written language or numerical figures, it can be image, text, audio, video that carrying people’s shopping, traveling, business, sports, entertainment and social behavior information, this heterogeneity and diversity is the embodiment of the ‘big’ data. In addition, 'Big' also embodies the relevance. Information of the library and the internet duration recorded by campus card must be associated with the students’ comprehensive achievements, real events happened within a few minutes will be related to the search engine and become a piece of news etc.. This characteristic has led directly to the study of the relevance of data variables gradually more than the causality study.

**Interpret ‘Big Data’ from the ‘Data’ Side**

From the perspective of performance or storage, big data contains 25% of structured data, 5% of semi-structured data and 70 of unstructured data.70% unstructured data (Imagination/Document/Audio/Video/Correspondence); 25% structured data (relational database, structured data file system/application/log data resides in data storage); 5% semi-structured data (email/XML).

From the perspective of data approaching, there are three kinds of approaching ways: first, generated by people communication and social networks, the second is through the man-machine dialogue and man-machine exchange, the third is the sensor data and machine data from objects to objects.

From the functional point of view, big data can be divided into transactional data, process data and interactive data.

In summary, from a statistical point of view, the concepts, characteristics and value of big data are mainly reflected in the timeliness, completeness, capacity expansion, type diversity, and dynamic accumulation.
Summary

From the perspective of statistics, ‘big data’ means population under certain conditions, and it can be extended, excavated and it also embodies the relevance and diversity.

Acknowledgement

This research was financially supported by the Chuxiong Normal University XJRC1603.

Reference