Sampling Survey among Undergraduates in the Age of Big Data

Chongjun Ouyang [a], Hu Wang [b]

School of Communication and Information, Beijing University of Posts and Telecommunications, Beijing 100876, China

[a]DragonAim@bupt.edu.cn, [b]AnyOneWho@bupt.edu.cn

Keywords: big data, sampling survey, undergraduates.

Abstract. With the development of information science, traditional sampling survey has been challenged and influenced by the trends of big data. Undergraduates are the most important foundation of the development of science and times. This paper analyzes the merits and faults of big data, and makes detailed analysis on the comparison between big data and sampling survey. Research indicate that big data has obvious merits than sampling but sampling survey is much more suitable for undergraduates' development, and undergraduates shouldn't just abandon sampling survey owing to the convenience of big data. Despite that, most undergraduates should just master the basic concept of big data and they should continue sampling survey in their innovations and investigations.

1. Introduction

College students must do researches when they want to do some innovations or creations. Also, when they have to tackle some social or scientific investigations, they have to turn to survey, too. As we all know, college students get used to sampling survey when faced with innovations or investigations for the limit of the time, ability. Recently, big data has proved its efficiency and power in the religion of business and politics by the great success it has obtained. In that way, people may advise that students change their minds, accept and apply the concept of "Big Data" in their innovations or investigations. Mayer-Schonberger, V. and Cukier, K. illustrate that sampling survey has no meaning of existence in the age of big data in their book [1]. This article tries to focus on whether undergraduates should abandon the sampling survey gradually and hug the big data; it points out the significance and feasibility of sampling survey to most undergraduates and it illustrates the relation between the big data and sampling survey among college students. The significance for this article is that most undergraduates just need to comprehend the concept of "Big Data" and establish the corresponding thinking mode; they should continue the sampling survey during their innovations or investigations because the way of" Big Data" is unpractical for most students.

2. The merits of big data

Big data marks an important step in humankind's quest to quantify and understand the world. There are innumerable things that couldn't be analyzed long before which have become measurable and analyzable thanks to big data. Mayer-Schonberger, V. and Cukier, K. have enumerated three shifts caused by big data in their book [1]. Zhi Geng and many other scholars have demonstrated the advantages of big data and the opportunity that big data could provide [2,3]. Basing these points, this article collects and synthesizes the merits of big data as followings.

2.1 The sample is equal to the overall

Big data tells us to process all of the data relating to a particular phenomenon. Big data is the results of a preponderance of things from the technology to the market demand even the government affairs. Once upon a time, we can't even store all the data relating to a phenomenon, not to talk about computing and processing it with the limit of central processing unit (CPU), which introduces the traditional sampling survey. However, this situation has developed greatly with the development of cloud computing and many other technologies, which permits us to store, compute and analyze all the relating data when doing some innovations and investigations.

Big data offers a thinking mode of dealing with all the data and gives birth to a revolution. At the point of the amount of processed data, it is different from sampling survey enormously. Because big data would process all the data, in other words, the sample is equal to the overall, we can not only broadcast the basic trend of the subject of our research but also catch a glimpse of the details in the issue. Google has ever identified the areas infected by the flu virus by what people searched for on the Internet. To do it instantaneously, Google took the 50 million most common search terms, which covers nearly all the relating data. And it works well and helps the government to control the flu virus. Nowadays, there are much more examples in which all the relating data is measured and analyzed basing the concept of big data. We can find the charisma of big data penetrating them.

2.2 More measurement error with less error from sampling

Big data accepts more measurement error with less error from sampling. When we analyze the issue with the method of sampling survey, we have to select each sampled individuality with great care because the data used is limitless and countless. In sampling survey, the data should be canonical and precise; but in big data, we needn't devote too many costs in sampling for the sample equaling to the overall. Big data aims at processing all the relating data corresponding to our research subject, so we needn't concentrate the accuracy of each items. It seems that this would lead to indeterminacy and error in the measurement, but it would gain reparation from the thorough sampling.

The superiority of quantity overall can make up to the inferior position of quality individually. So, it may not reveal the shining of big data because it seems like not an improvement of results but a transfer of error sources. But it should be realized that too much time, money and many other costs would be throw in the guarantee of sampling individual, which would lower the timeliness or immediacy greatly when analyzing the problem or issue. When customers go through the shopping website like Amazon.com or Taobao.com, their choices and preference may change all of a sudden. And with the trend of fashion, the popular and the acclaimed would chop and change. So the recommended goods or cargos should transform correspondingly, which requires excellent immediacy. The sampling survey couldn't reach the requirement and this would be solved satisfactorily by big data for its obvious strengths.

2.3 About what not why

Big data doesn't concentrate on causality like age-old search any more, it focuses on the patterns and correlations in the data. Briefly, big data is about what not why. In sampling survey, the most precise and typical data would be collected and analyzed basing kinds of statistical methods. Some methods have been existing for thousands of years, and the time has proved its practicability; sampling survey could reveal a definite causality when analyzing and exploring some issues. However, most things can't be measured by a definite theory or equation because things would change with its environment and some inherent factors. In the age of big data, all the relating data would be measured and analyzed without selection and correction, so the result may not be a definite causality but a probability.

In sampling survey, the result for the analysis may be illustrated that A is caused by B. But big data would not only show its trend but also its internal details, so the results may be there is p% for A_1 happening, q% for A_2 happening by the constraints of B. In many cases, the latter illustration is much more sophisticated. In some consulting mechanisms, when they help others to make decisions or provide advice, they would need a comprehensive solution for the consulting. And big data would handle this quite well, which would come up with an overall and detailed scheme. Also, in a security company or a stock and share company, the scheme with probability would be much more efficient than the scheme with specific causality, as the stock changes like the lightening.

2.4 Summary

Basing the illustration above, the strengths of big data is transparent. It can ensure the high immediacy, high accuracy and high comprehensiveness. So undergraduates must shape their concept and thinking mode of big data.

3. The faults of big data for undergraduates

We have demonstrated the merits of big data above. Basing the discussion, there is no doubt that big data has apparent strengths in statistics. However, big data has some intrinsic faults companying itself along with the advantages. Ying Wang and many other scholars have demonstrated the faults or some questions of big data [4,5,6]. Most of their statements focus on the strengths of big data, and then try to exemplify the inherent problems hiding in its virtues. As this article focus on the undergraduates, so the statement would focus on the faults of big data for undergraduates.

3.1 High expectations for techniques

Big data is the result of technology, market demand and many other factors. But the most essential thing is the technology. Without fantastic technology and effective scientific method, all the conjecture about big data would go bankruptcy. Big data acquires the fundamental of data mining, distributed computing, the corresponding knowledge of cloud computing and high-authority algorithms used to process the data. All of these techniques are on the basis of familiarity with data science and computer science.

There are countless undergraduates all around the world, who major in different disciplines and have great disparities in data processing and computing. Most students even couldn't get to understand the fundamental concepts of these techniques all their college careers not to mention relying them to do big data processing. Only a few of the undergraduates with great talents or interests or specific majors could practice these skills roughly, but it's still hard for them to abandon sampling survey when do their innovation or investigation. In this way, we can seize that big data show great challenges in techniques or technologies, and the techniques are too obscure for most undergraduates.

Undergraduates should be encouraged to do innovations or investigations, but the innovations should correspond to its power and levels. Bid data is such a sharp weapon in scientific exploration, but it requires a highly profound ability far more than conventional undergraduates. If a student majoring in history want to do some researches about the sword, he or she must collect as much the relating data or documents as possible. So, the most appropriate method is to do data mining on the Internet for the data or documents has been stored in digital ways rather than analogy ways. And then he or she can use the corresponding technologies about big data to do some measurement or analyzing on this data. But he or she is a liberal art student, the requirement for techniques would be hard for him or her. At last, he or she may abandon this investigation upon swords, which has deviated from his or her origin.

On the whole, big data challenges undergraduates much more sharply in techniques and methods than the sampling survey.

3.2 Different and misplaced targets

We have to demonstrate again that big data is the birth of technologies and market demand. Google used big data to broadcast the virus flu precisely and immediately. Amazon.com uses big data to introduce new books or other goods to its scanners. Oren Etzioni used big data to broadcast the plane tickets before travelling or hanging out. Donald Trump used big data to win his election and gain public's focusing. In this way, we can find that most successful cases about big data are involved in politics or finance because they require high immediately and accuracy. The three main strengths may seem prepared for their use originally. We don't mean that big data shouldn't be introduced in scientific or social investigations or researches. But we can see from these examples that big data may not be appropriate for undergraduates' using because their targets are distinct from each other.

Big data don't concentrate on the accuracy and rationality of sampling items and it aims to accepting more measurement error with less error from sampling. But for undergraduates, the fundamental of technologies in big data may be hard and untouchable for them; these abilities can be formed and seized in their master's careers. But the basic skills or virtues about the sampling survey must be acquired and proficient during their college lives. They must learn to cherish the sampling data or the experimental results; also, they must obtain the basic ability of discriminating. They should learn how to store, measure and even analyze data, and they must learn to abandon inconsequent results. That's the basic point of scientific exploration they should cultivate as undergraduates.

Big data don't care about the quality of each sampling point and it can't develop undergraduates' scientific spirits. This could be regarded as the second fault of big data to undergraduates. Big data has a different purpose and methods from basic educational points. Students should be advised to gain the basic spirits from the innovations and investigations, that's the cause of big data's fault.

3.3 Neglect of causality

In the statement about the merits of big data, we have pointed out that big data doesn't concentrate on causality like age-old search any more, it focusses on the patterns and correlations in the data. Though it's a fantastic strength for big data, but it will become the boundary for undergraduates. Most undergraduates want to do or taste innovation or investigation, also, sometimes their college will distribute some innovation tasks for them. To finish the innovation or investigation, they have to analyze the question they have to face and find out the realistic demand for their innovation. But the most important destination is to exercise their ability of finding or exploring causality. However, big data would destroy these basic points once it replaced the sampling survey.

For undergraduates, the innovation and the investigation are set to help them form the power of exploring using the fundamental knowledge they obtain in their college. And the process shouldn't stride over a long stage. The big data must take undergraduates much time and vigor for the limitless power and thinking mode, which would mean a boring and tedious process. And the skills for seizing causality is uneasy for undergraduates not to mention the probability analysis. So, we can see that the third strengths would become chain for undergraduates, too.

It should be emphasized that we don't mean the big data has four main faults in every field, we just mean the shortages for undergraduates. Most undergraduates get used to analyze the causality for most things, and they would apply this method in their innovations or investigations. Big data would just block their way to the final success although it's such an excellent method. If there is a student majoring in telecommunication engineer, and he wants to test the property of their campus network. He would then figure out some indexes to calculate the property of the network and collect data to explore them. As an undergraduate, he hopes to figure out some conclusions after some inquiry and researches. So, some definite causality is much more realistic than probability.

3.4 High economic cost

We have stated that big data requires many intensive techniques involved in data science and computer science. To process data with big data, undergraduates may have to contact with cloud computing, distributed computing and data mining. Data mining is used to store digital data from the internet; cloud computing and distributed computing are used to measure and analyze data. These techniques require not only high-authority algorithms but also superior hardware configuration. Although the cost of hardware has decreased a lot during recent years, but the fine hardware is a little expensive for undergraduates without incomes.

Some undergraduates may submit their innovation schemes or investigation plans to their college, and the colleges may offer some help both in fund and techniques. But that's not enough. Even the same algorithm would present different effects, the time their machines spend to execute their algorithm would vary a lot between each other. We can assume that there is a student who want to acquire enough data from the blogs to do some jobs on the analysis of social network. He or she may program some algorithms and carry out it on his or her computer. If the computing power of his computer is a little weak, it may take him months to get this data. But if his hardware configuration is superior, it may just take hours.

Basing the illustrations above, we can conclude that the big data would indicate largely high penalty for undergraduates. They may get some help from their colleges sometimes, but it would be troublesome for them when they want to do some innovations and investigations freely.

We have demonstrated all the faults of big data for undergraduates by now. From our statements, we can find big data is not satisfied for most undergraduates although it has some obvious strengths. Compared with big data, sampling survey is much more realistic and touchable for undergraduates. Although it has many shortages, it could teach the undergraduates scientific methods and even spirits.

4. Comparison between sampling survey and big data for undergraduates

We have identified that the merits and faults of big data. Basing the analysis above, we can conclude that big data has great superiorities than sampling survey, but it has some faults naturally which couldn't help or exercise undergraduates. It's important for undergraduates to cultivate the concept of big data, but it's hard, unrealistic and no helpful for students to analyze issues in the way of big data. By the way. They should continue learning and mastering the sampling survey which would both challenge and exercise them appropriately. There are many researches and papers talking about the merits of sampling survey especially in the age of big data. Wenfang Tang and many other scholars or agencies have done some outstanding in the significance sampling survey at the age of big data [7,8]. The comparisons between sampling survey and big data for undergraduates would be illustrated in the followings.

4.1 The sampling survey can make up to the faults of big data ideally

In the statement of the faults of big data, we have declared four main faults of big data for undergraduates, which makes it hard and unrealistic to do analysis with big data. Compared with big data, sampling survey could solve these faults or shorts well. Big data requires a deal when doing innovations and investigations both for time and hardware configuration. Sampling survey also requires time, but the money or fund spent on it would be much lower than big data. The cheapest method to collect data and analyze it is to use the questionnaires. With the spread of Internet, the dispensing of the questionnaires would be much convenient than the old times. Sampling survey plays an important and crucial role in cutting the costs in innovation and investigations for undergraduates. Also, sampling survey doesn't require too many specialize techniques and tools. After gathering information from the questionnaires, students may filter the reasonable and appropriate items for their projects, and then they would turn to the traditional statistics for analysis. This is different from big data, because students needn't master the professional knowledge of data science or computer science. It makes innovations and investigations much more feasible and near for themselves. If a team want to finish an investigation about the spread of Internet in the villages, they can design some questionnaires and hand out them in the villages they have picked up. The costs on the traffic in sampling survey would be much lower than the costs on the hardware, algorithms and time in big data. It should be emphasized that undergraduates should keep on sampling survey in their innovation and investigations, and they should just master the concept and thinking mode of big data. Maybe big data would require a much lower costs for undergraduates after years, but it's hard and unrealistic for undergraduates by now.

4.2 The sampling survey could inherit the basic scientific spirits

For undergraduates, the essential lesson that should be taught in the college is the scientific spirits. And they must learn enough skills to match these spirits. For big data, we have declared that it couldn't teach students the basic skills and natural spirits for its nature. In the statement of the merits and faults of big data, we have demonstrated that big data don't care about the accuracy of sampling items and the causality. But for undergraduates, it's significant to learn how to filter the reasonable and appropriate data from the sampling results. Considering the ability of them, it is much more efficient to explore the causality instead of the probability. Big data requires quantities but neglect the qualities, which is opposed to the education of undergraduates. Students come to college and they may become excellent scientific scholars or social scholars in the future. They have to exercise their ability of differentiate infallibility and analyze errors, and they should have the accomplishment to explore the causality rather than the probability, which would confuse them instead of helping them. But sampling survey would offset the shortages of big data. In sampling survey, undergraduates would spend much time on testing the sampling results to get the final samples with high accuracy. After selecting the sample points, they would process the data with traditional statistics which would lead them to an exploration of causality. We can see from sampling survey that it could exercise and help undergraduates at the stage of undergraduates. These skills and experiences offered by sampling survey would help college students practice scientific spirits and apply basic analysis methods, which would be the basis of big data in their future career. Big data is such a sharp weapon that it may act link a “black box” and it neglects many important details in analyzing issues. It may be much more convenient for this neglect in practical problems but these details are just the goals undergraduates hope to achieve by innovations and investigations.

4.3 Discriminate between sampling survey and big data

Basing all the statement above, we can now conclude that though sampling survey has many shortages, it is the most appropriate method for the undergraduates. There three main merits of big data along with four main faults or shortages for undergraduates. We should emphasize again that we don’t mean the big data has four main faults in every field, we just mean the shortages for undergraduates. Big data has many advantages but it’s not good for students’ education and the formation of scientific spirits. Sampling survey may not show the details of some issues except the trend and it analyzes causality without probability. It seems that it’s such a not all-inclusive method and it’s a little arbitrary. But with sampling survey, the ability of analyzing the accuracy of data and causality would enhance greatly. Big data don’t care about the accuracy of sampling points, which leads to more measurement error with less error from sampling. In that way, undergraduates won’t cherish the experimental results anymore, because it would be blamed to the nature of big data. However, experimental results would be the key point for undergraduates, that’s the essence for their innovations and investigations, which could help to develop a rigorous scientific attitude. Assuming that there two groups of undergraduates who are exploring the same question, one group chooses the method of sampling survey, the other turns to big data. The group in sampling survey would design some questionnaires and distribute them, and they would soon finish the sampling and analyzing. At last, they would come to a definite causality. But for the other one, they would spend too much time on data mining and algorithms designing. At last, they would come to a conclusion with probability. If this persists, the first group would have the courage and interests to accept more investigations and make some innovations at last. But for the other group, they would find it tedious and boring, they may lose their patience and abandon innovations. Undergraduates need encouragement and help, so we can see from this example that sampling survey may be much more suitable for undergraduates.

Basing the statements about the comparisons between big data and sampling survey for undergraduates, we can realize the merits and faults of big data again. But after the comparisons, we would treat the big data with more caution. For undergraduates, they may get used to sampling survey not only for the old tradition but for the realism and practicality. This would be much helpful for the development of students. But big data is a two-edged sword which is sharp for the analyzing but it may hurt the undergraduates. It is no use in education of traditional scientific spirits, and it won’t do well to the development of most undergraduates.

5. Conclusion

The human history has come through the ‘Information Age’, which gives birth to big data. And more and more scholars and agencies even government turn to analyze issues with big data, which challenges the traditional sampling survey. This article focuses on the undergraduates under the age of big data, and explores whether they should abandon sampling survey and apply big data into their daily innovations and investigations. The merits and faults have been stated and demonstrated overall in this research. Despite that big data has great advantages, most undergraduates shouldn’t abandon sampling survey because it’s essentially typical for their development. Big data has its own advantages, but sampling survey basing “small data” has its advantages, too. But it should be emphasized that big data is not suitable for the development and cultivation of undergraduates’ spirits and basic skills. The conclusion comes that most should just master the concept and thinking mode but they should continue applying sampling survey into their innovations and investigations except some students majoring in corresponding discipline. This conclusion is obtained by the limit of costs of big data. Once big data become much more common than now, the conclusion should be adjusted. And big data could be applied to students’ innovation and investigations.

References

[1]. Viktor Mayer-Schonberger, Kenneth Cook. Big Data a revolution: that will transform how we live, work and think. Zhejiang people's publishing house, 2013, p. 27-45

[2]. Zhi Geng. Opportunity and Challenges of Big Data to Statistics. Statistical Research. Vol. 31 (2014) No. 1, p. 5-9

[3]. Hequan Wu. Opportunities and Challenges in the Era of Big Data. Chinese Storage & Transport. Vol. 3 (2013), p. 53-54

[4]. Ying Wang, Shuchen Wan. Challenges and Opportunities of Sampling Survey in the Age of Big Data. Statistics & Information Forum. Vol. 31 (2016) No. 6, p. 33-36

[5]. Kenneth Neil Churchill. Challenges and Limitations of Big Data. New Marketing. Vol. 9 (2013), p.12-12

[6]. Tao Ma. A Rational Perspective on Big Data. Information & Communication Technology. Vol. 6 (2013), p.58-62

[7]. Wenfang Tang. Big Data and Small Data: A Reflection on Social Science Research Methodology. Journal of Sun Yat-Sen University (Social Science Edition), Vol. 55 (2015) No. 060, p.141-146

[8]. Liming Pan. A Brief Account of the Significance of Sampling Survey in the Era of Big Data. Modern Economic Information,Vol. 1 (2016), p.10