Study on the Impact of Big Data on Radio and Television Ratings

Chai Jianping[1, a], Pan Xingyi[2,b], Yin Fulian[3,c] and Chai Juanfang[4,d]

[1,2,3]Communication University of China, Beijing, China

[4]Shanghai Electromechanical Engineering Institute, Shanghai, China

[a]jp_chai@sina.com, [b]penny_pxy@hotmail.com, [c]yinfulian@cuc.edu.cn, [d]chaijuanfang2006@163.com

Keywords: Radio, Television, Big Data, Ratings, Sampling, Full Sample

Abstract. For problems, such as the low accuracy and easily polluted data of the traditional sampling ratings index calculation method, we conduct research on the ratings of big data. Firstly, we argue the effect of the amount of data on the accuracy of the ratings, and do experiments using real data. By sampling with replacement, sampling without replacement and stratified sampling, we verify that the increase of the amount of data will improve the accuracy of the estimation. Finally, this paper explains the necessity of radio and television big data analysis.

Introduction

The Big Data era has arrived, the data has penetrated into all walks of life, becoming an important production factor. At a general sense, big data refers to data collection that cannot be perceived, acquired, managed, processed or served by traditional IT technology, hardware and software tools within a tolerable time[1].Broadcasting industry in terms of big data has a natural advantage, which itself is a sector with a large number of data resources. The recent ESG report shows that, compared to other industries' 24% annual growth rate of the data, the annual growth rate of data media and film industry is 31%. On average, the companies in this field need to use 30% of the overall IT budget for data storage and the proportion is far higher than other industries [9]. The broadcasting industry which has a massive data resources must step up the pace, using their own advantages to catch up with the era of big data, otherwise it will be abandoned.

There are many flaws with ratings obtained using traditional sampling methods. In order to overcome the drawbacks of traditional sampling methods, domestic media industry are beginning to use big data technology to change the way data is collected. CCTV Sofres (CSM) can monitor the ratings of more than 1000 major television channels from 23 provinces and 124 cities 24/7. October 23, 2012, Blockbuster New Media Institute in Shanghai announced a new media viewership evaluation system - "Blockbuster new media content index"[3].The agency use the "all media coverage data research" system which has independent intellectual property rights to get the massive interaction data from all collected tens of millions of users. By analyzing and precisely calculating the power-frequency, length and other data when viewing, we formed a scientific, rigorous evaluation system. March 20, 2014, Ali cloud announced joint CDV, China-cloud data, to build China's largest all-media cloud computing platform. The platform may help traditional television network into a multi-screen television, support computer website, mobile APP, TV full terminal smooth playback, and can collect and operate the big data [4]. American television company Nielsen in September 2013 equipped his family's 23,000 samples with a new system which can include the Internet viewing content into the ratings statistical range. This system combines the streaming data and television ratings data together. It tracks not only the Amazon, Netflix, Hulu and online video-on-demand situation on the Xbox, but also the viewership iPad platform. The American famous online movie rental provider Netflix, will analyze the 30 million plays, 4,000,000 ratings, three million searches and the equipment used by the registered users. To this end, Netflix's television programs are labeled hundreds of labels [5]. Data mining of large companies determine the understanding of audience preferences Netflix has far exceeded its competitors.

For problems, such as the low accuracy and easily polluted data of the traditional sampling ratings index calculation method, this paper carried out a calculation based on the ratings of big data research.

Theoretical Analysis

The purpose of a sample survey is to use the sample survey data to infer the whole.But it is impossible to get the exact total true value from any single sample. The error which exists in the survey data is absolute, while the size of the error is relative. We need the index relative error to be little, the data accuracy to be high and so the demand for the sample size is large.In order to theoretically prove that "the amount of data, the higher the precision", this paper uses the relevant knowledge of statistics and three methods to demonstrate.The data used are the all 2,187,648 ratings data in one day from a Chinese province, covering user ID, channel ID, watch the start and end time.

Method 1: use a confidence interval.In large sample, we usually use the formula:

$$\pi = P \pm 1.96 \times \sqrt{\frac{P(1-P)}{n}} \quad . \tag{1}$$

to calculating the confidence interval.

In Eq. 1, P is the sample proportion, n is the sample size, π is the overall proportion, taking 95 percent confidence level. ,for example, we use samples with replacement to calculate the arrival rate. When the number of our sample is 85% of the province, that is 1,859,500 people, the arrival rate is 10.77%, and $\pi = 0.1077 \pm 0.00045$. Increasing the sample size n, we obtain the following table:

Table 1. theoretical confidence intervals（sampling with replacement）

n	1859500	1900000	1950000
π	0.1077±0.000450	0.1077±0.000440	0.1077±0.000435
n	2000000	2500000	3000000
π	0.1077±0.000430	0.1077±0.000380	0.1077±0.000350

From the Table 1, it can be seen, as the sample size n increases, the confidence intervals become narrower, the data are more accurate, which is intuitively correct, because larger sample contains more information, and therefore have a more precise conclusions.

Method 2: Using the relative error argument. Relative error formula is:

$$rp = 1.96 \times \sqrt{\frac{p(1-p)}{n} / p} \quad . \tag{2}$$

In Eq. 2 P is viewing index, n is the sample size, the degree of confidence t is 95%. for example, use samples with replacement to calculate the ratings, we obtain the following table:

Table 2. Ratings theoretical relative error（sampling with replacement）

Sample size	Ratings	Theoretical relative error
15%	0.167%	9.945%
30%	0.309%	5.166%
35%	0.339%	4.566%
50%	0.460%	3.277%
55%	0.522%	2.938%
80%	0.738%	2.043%
85%	0.774%	1.935%
90%	0.819%	1.827%
95%	0.857%	1.738%

As can be seen from the table, with the increase in sample size, calculated by a formula theoretical relative error continues to decline.

Method 3: use the reverse thinking, calculating how much amount of data is needed when the absolute error is limited to a certain range. Apply the formula to calculate the sample size large sample size calculations under different absolute error:

$$n = \left(\frac{t}{2 \times \Delta p} \right)^2 . \tag{2}$$

In Eq. 3 t is the confidence level and Δp is the absolute error. We obtain the following table.

Table 3. Sample Size

Confidencet1	Confidencet2	Confidencet3	Absolute error Δp	Sample size n1	Sample size n2	Sample size n3
1.65	1.96	2.33	2.00%	1702	2401	3387
1.65	1.96	2.33	1.00%	6806	9603	13544
1.65	1.96	2.33	0.5%	27225	38414	54178
1.65	1.96	2.33	0.3%	75625	106707	150494
1.65	1.96	2.33	0.15%	302500	426828	601975
1.65	1.96	2.33	0.075%	1210000	1707378	2407900

Table 3 shows that the smaller the absolute error is, the greater the confidence level and the sample size.

By the three methods above, we can conclude that: increasing the amount of data will help to reduce errors, when you want to limit the error within a small range, but also requires a lot of data.

Experimental Results Analysis

The broadcasting industry has its incomparable advantages, but television viewing market has seen a continued contraction in nature worldwide. This crisis is caused by many reasons. First, In recent years, television services which can provide similar types of business have increased rapidly; Secondly, the traditional viewing index system and traditional sampling methods can not precisely comment the influence and quality of the television programs. Fake ratings leads to the result that the they can not reflect the quality and impact of the program.

For the problems of the traditional way, we use the Differentiation sampling experiments to analyze. The data used are all 2,187,648 ratings data in one day in a Chinese province, covering user ID, channel ID, watch the start and end time.

Sampling with replacement. It means that every time you extract a sample at a certain probability, and put it back after extraction, and then the next sample, each sample are independent. Respectively extract 10%, 15%, 30%, 35%, 50%, 55%, 80%, 85%, 90%, 95% to calculate viewing indicators, and compare them with the ones which use the data of the full province.

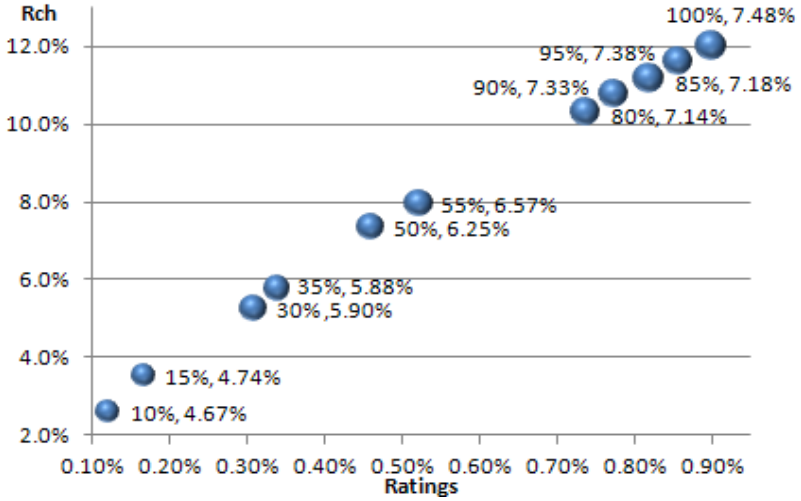

Figure1. Sample index ratings with replacement

In Fig.1 the horizontal axis is the ratings and the vertical axis is the arrival rate, blistering size represents loyalty. As can be seen, with the increase in the number of samples, three indicators are rising, and gradually approaching the data calculated using the results of the province.

Sampling without replacement. It means to extract samples in accordance with a certain probability, and continue to take the remaining data at a certain probability, lack of independence. Separately take 218765,334711,441227,536826,656119 to calculate viewing indicators, and compare them with the ones which use the data of the full province.

In Fig. 2, the horizontal axis is the ratings and the vertical axis is the arrival rate, blistering size indicates the number of ratings per minute. As can be seen from the figure, the use of the results of sampling without replacement was not satisfactory, although with increasing sample, three indicators are increasing, but there are significant differences with the results of the full sample data.

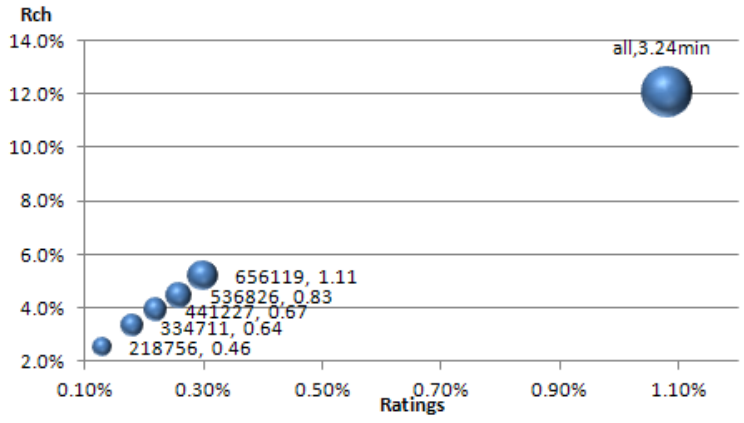

Figure2. Sampling without replacement indicator ratings

Stratified sampling. It means firstly to put the overall unit by certain characteristics into several sub-population (layer), and then do the simple random sampling from within each layer. Extract data according to different cities of the province as 10%, 30%, 50%, 80%, and then gather the extracted data.

In Fig. 3, the horizontal axis is the arrival rate and the vertical axis is the number of ratings per minute, blistering size represents loyalty. As can be seen from the figure, with the increase of the

sample, three indicators are rising, and gradually approaching the province's data with the calculated results.

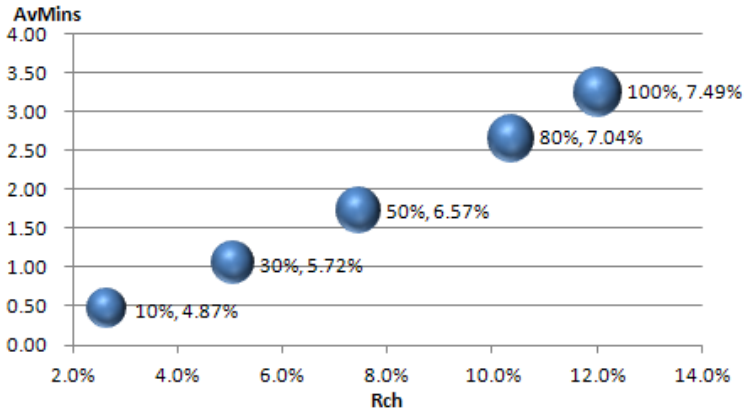

Figure3. Stratified sampling with replacement indicator ratings

As can be seen from the three experiments, when using sampling with replacement, sampling without replacement and stratified sampling with the number of samples increases, the ratings index value will continue to increase, and continue to approximate the results of a complete sample of the province, and stabilized. The main reason for this phenomenon will be analyzed as follows:

First of all, no matter which way we take,we extract part of the data . When the amount of data is small, an accidental behavior is not representative is likely to affect the results. However, with the increasing amount of data, even if some dirty data mixed in, they will be inundated vast amounts of data, the impact on the final result will be gradually reduced.

Second, when we extract a small sample of data to calculate the index ratings, we may not extract the representative households. For example, the sample may contain the results of a large number of users interested in a program A, which causes the ratings of A to be high. In addition, the burst will affect the user's viewing behavior ratings index analysis, but these acts are often not representative. For example, viewing behavior of some part of the sample households are concentrated in a particular day or days, and the results will be biased.

In addition, different age groups, income, level of education will affect the user's viewing behavior. For example, women may be more like to watch cooking shows, while children like cartoons. Men might prefer news programs. If we just extracted part of the sample households, it is difficult to cover all sectors of the population, then the result will be biased. However, if you are using big data, which has enough information, you can get a more reliable ratings trend.

All in all, the small sample has its inherent flaws, it does not have random, unable to effectively resist the sudden unusual circumstances, can not cover all sectors, which will lead to the final inaccurate result.

Necessity of Radio and Television Big Data Analysis

Traditional sampling is bound to be abandoned era, the era of big data has arrived, and now the "big data" are much more discussed in the broadcasting industry, so in the end what is useful for the broadcasting industry.

Using big data analysis, inevitably there will be some dirty data incorporated. For example, the set-top box is not closed all night, but the user does not have any viewing behavior. If using traditional sampling methods, this is not representative and will result in a great impact on small sample data. Theoretically such data should be removed, but in fact, there is no way to effectively eliminate erroneous data. In the vast amounts of data, this error is no longer evident.

Before the era of big data, people need to establish a sound model, design sophisticated algorithms to solve the problem. Thus, the question depends on whether the model is reasonable and whether the algorithm is efficient. The big data are better than algorithms in many areas. Increasing the amount of

data will help greatly improve the algorithm's performance. In this paper, a method of increasing the amount of data in the ratings, the arrival rate and per capita ratings minutes algorithms in certain circumstances. No matter in what methods, with the increasing amount of data, and finally the data obtained are steadily and gradually approaching the true result. Broadcasting industry itself has a lot of data, the broadcasting industry can take advantage of their strengths, with the vast amounts of data to ensure the accuracy of the data analysis results.

Full sample data are more complete than the sampling data, and more comprehensive. According to the city sampling, which is to use the full sample calculation of a city, and then increase the full sample of a city finally to the province, you will find a very interesting phenomenon: With the increase in the number of cities , the ratings index is also growing, but the growth rate is far better than other sampling methods, and each sample being counted out of the ratings indicators and the true value is not far off. The results are shown in Figure 4-1. The results demonstrate the superiority of big data: Even if there is only one city, we get all the data, it is a complete sample of the user structure,. The calculated indicators are nearly the same as the ones of the whole province, that is to say a complete data are more persuasive and stable than conventional sampling index calculated ratings.

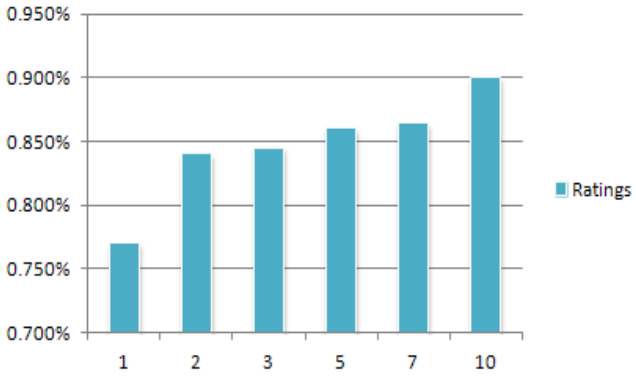

Figure4. Ratings (by city sampling)

To step into the era of big data as soon as possible, the broadcasting industry must make a change:

First, change of data collection. If the broadcasting industry in the country can use the two-way network set-top boxes automatically return the way to collect the massive data from the entire network, then you can avoid the sample bias, and a series of sample contamination problems which will affect the results of the analysis, thus greatly improving the ratings index reliability.

Secondly, the increase in the number and types of data. Today's data sent back by the set-top box only indicates the program, start and end time and other basic information, there are some limitations based on the results of these data drawn from the analysis. But if the program makers can obtain the user's demographic information, such as age, education level, occupation, we can help show producer targeted tailored to different people of different programs, and to arrange for playback at a reasonable time, to achieve maximum benefits.

Third, improve the ratings indicator system. Today's media environment has changed dramatically, video users are no longer just in front of the television viewers, radio and television industry, the ratings must now make improvements as soon as possible indicators can no longer rely solely on traditional viewership ratings and other indicators. The industry needs to be on the Internet user's click-through rate, user evaluations and opinions into account areas.

Summary

Big data trend is inevitable, radio and television industry needs to accelerate the process of triple play, breaking data silos, to cooperate actively with the Internet. Firstly, in terms of confidence intervals and relative errors, the paper proved theoretically that increasing the amount of data helps to reduce errors. Then it uses reverse thinking from the perspective of the sample size, the results show that when we hope the error to be in a very small range, it requires a lot of data. Then this paper uses

actual data on the ratings index calculation. The experimental results fully illustrate that the small sample has its inherent flaws, it does not have random, unable to effectively resist the sudden unusual circumstances and can not cover all sectors, which will lead to the final inaccurate result. Finally, the paper discusses the necessity of using big data in the broadcasting industry and gives some suggestions. In short, the task of the media is a long way to go, I hope the broadcasting industry can make good use of big data, produce excellent programs to achieve maximum benefits; hope to be able to take advantage of the triple play of the broadcasting industry opportunities, grasp the initiative to promote their better development.

References

[1] Li Guojie, Cheng Xueqi. Research status and scientific thinking of big data. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647-657(in Chiese)

[2] The 2011 Digital Universe Study: Extracting Value from Chaos. International Data Corporation and EMC, June 2011.

[3] Blockbuster: the establishment of new media viewership evaluation system [J] Chinese broadcasts,2013,01:80.

[4] Ali cloud, cloud data Watertown, New Auto jointly build China's largest all-media cloud platform [J]. TV works,2014,01:2.

[5] Changes in the era of big data to the broadcasting industry to bring Yan Qing. [J]. China Digital TV,2013,Z3:37-38.

[6] Wang Jianlei big change in the nature of the data gives the broadcasting industry [J]. Sound Screen World,2013,09:8-11.

[7] Victor Meyer. Schonberg Big Data era [M]. Zhejiang People's Publishing House, 2012-12.

[8] Du Zhang. Inconsistencies in big data[J]. Cognitive Informatics & Cognitive Computing (ICCI*CC).2013.

[9] In the Greek National: "High-definition and big data management influence in the film and television post-production and broadcast of "http://tech.ccidnet.com/art/40939/20130705/5052869_1.html (2013/7/5)

[10] Zheng Weidong. Ratings and big data [J]. Advertising Grand (Comprehensive Edition),2014,05:124.

[11] Under Ma Tianshu broadcasting industry. On the background of big data [J]. Cable Television Technology,2013,04:126-128.

[12] The correspondents. Big Data era, the media how to transition [J]. Friends Editors,2013,06:6-12.

[13] Pan Hongtao. Rethinking ratings assessment system under the background of big data [J]. Chinese Journal of Radio and Television Studies,2013,07:17-19.

[14] Zhou Yunqian television in response to the changing situation and the road Big Data era [J]. Chinese TV,2013,09:90-93.

[15] Xu Qi Innovation Evolution Big Data era ratings survey [J]. Media Observation,2013,10:28-30.

[16] She Xianjun data fog of media literacy [J]. Chinese advertising,2014,04:138-139.