An X-method to Get Sensitive Information

Liu Xiang
Math Department, Science College,Tong ji University
Shanghai,200092 China
fashengliu@163.com

Abstract—A new approach, which is so called, to get the sensitive information is presented. The X-method consummates and improves the randomized response technique (RRT). It makes up essentially the limitation of Warner and Simmons models. It confirms both the deference of the individuals and the accuracy of the information. The result can be used in various sensitive information survey, such as the total number of the cheat-ing students in examination, under table income, tax dodg-ing, and ladies age, etc. It has potential applications in designing electronic device systems for information processing.

Keywords : Sensitive information; randomized response technique; Warner model; X-method.

1. INTRODUCTION

There are various sensitive and private information survey issues in the modern society. The sensitive information is about that not easy to show publicly, such as if a company steals tax or not, and how much if it does. In a school how many students are cheating in an examination? The number of people of drug taking and gambling, the aver age saving of the people, the volume of smuggling, the recorder of crime, the homosexual activity, etc. and all the related information are sensitive information. It is a contradiction to protect the private information but also to socialize the information. How to deal with the contradiction perfectly is refer to both social science and mathematics. It is also refer to physical model and electronic device. The electronic lie detector is an extreme production for the issue.

It is hard for people to believe the result. It won't protect the privacy, but also has suspicion to overpass human right. Therefore it is difficult to apply it widely. It is necessary to design a reasonable sensitive information surveying method. To develop a method both protecting privacy and guaranteeing accuracy has important theoretical and social practice meaning.

We are going to give a new approach to get the sensitive information without damaging the privacy and the accuracy of information by algebra method and a physical device.

People are usually very curious on social sensitive information. From main reason we do not really want to spy upon someone's problems that hard to open his or her mouth. We just want to know the total number of group or percentage of the people about some sensitive information. These data information is crucial in social decision area of many important fields. But people usually have more or less vanity or guarding mentality and won't provide their sensitive information publicly. Even like the objective data of age that can hard be gotten easily in 100% accuracy. How can we talk about the real sensitive question? For example, many people concealed the number of their children in China's fifth population lustrum because they are afraid of being punished. The person who breaks the rule of family planning is afraid of saying the truth. In fact the government just wants the real data to govern the society better. One side the investigator has to deference the

978-1-4244-8694-6/11/$26.00 ©2011 IEEE

object, on the other hand total information is required accurately. It's a contradiction to protect the privacy and to keep the data accuracy in such information survey pro- grams. The following methods can help us to get the sensitive information.

2 . SENSITIVE QUESTION AND ITS RESPONSE TECHNIQUE

In 1965, Mr. Warner put forward the RRT, which is called Warner model by the followers. It started the RRT for sensitive information issues. Its design principle is to have two contrary questions by the sensitive issue. The objects will choose one answer in a proper probability. The investigator has no right to ask the object which question he or she answered. In this way the secret is protected.

The sampling method is widely used in various in- formation survey situations. It is easy and effective for its simplicity. The basic data must be authentic in order to get the result believable. For sensitive information, it is hard to get the fact directly if we can't say it is impossible. So it is desirable to use the Randomized Response Technique (RRT) to get the sensitive data.

RRT has been paid more attentions since Warner presented the so-called Warner model in 1965. It provides a device in order to let the objects to answer the sensitive question in a probability instead of answer it directly. For example, to investigate the number of students cheating in examination, the investigator designing N cards, among them M cards with the question " Did you cheat in the exam?" and the contrary questions on the rest N-M cards. The objects answer the question according to the cards and their real situation independently and the investigator has no right to know which kind question except the answer itself. The private information for the objects has been kept in somewhat and they are willing to cooperate with the survey program.

Sensitive question can be classified into two categories. One is earmark or attribute sensitive information problem and the other is quantity sensitive information problem. Form the binary information system they are equivalent to each other. Here we just introduce the earmark sensitive information model and its estimation for the parameter.

Suppose there are two non-joint classes S and in a group, and there is p percentage of persons belongs to S and q percentage of cards with question "Are you in S?" We want to estimate the probability p of S.

Under the simple random sampling with redrawing situation (SRSWR), a sample of n has been got. We can give the estimation of pin the following way.

Let

$x_i = 1$, if answer is 'yes',

$x_i = 0$, if answer is 'no',

then

P (x_i= 1) = pq + (1 − p)(1 − q), i = 1, 2, . . . , n;

P (x_i= 0) = (1 − p)q + p(1 − q), i = 1, 2, . . . , n.

Suppose y persons answer 'yes' and n − y answer 'no'. The desirable estimation of p is

$\bar{p}$ = y/n − (1 − q)/ (2q − 1), (q = 1/2)

The method has shortcomings that we can't get the accurate information and the objects' privacy can't be guaranteed. Especially if q is about 0.5 we cannot get any information while if q is far from0.5 the plan has no privacy at all.

In 1967 Simmons improved the above model by introducing more independent questions rather than contrary ones. But it still can't solve the contra- diction between the privacy and accuracy.

3. THE LARGE NUMBER COVERING METHOD FOR SENSITIVE INFORMATION

To overcome the mentioned shortcomings, we use a large number to cover the sensitivity instead of l is as using RRT. Suppose we want to get the accurate number x with sensitive property in a group of N. We introduce an original large number x0 and let x k, (k = 1, . . . , n)be the k-th private data, we can get the x by summing the individual data one by one and deducting the original data. The mathematical mode follows.

$$x_i = 0$$

$$x = N - x_0$$

In the procedure, the k-th person only knows the information of $\sum_{k=0}^{k-1} x_k = s_k$ Using this method, the average under table income can be easily got dividing by the total number of objects. Of course if the (k + 1)-th and (k − 1)-th objects are discussing, the k th object has no privacy at all.

We can improve the method by introducing distribution for the information of x_k . That is to use vector to replace number x_k , where X_K has a distribution of n points with the algebra sum of x_k , The corresponding mathematical model is as follows.

$$\sum_{k=0}^{n} X_k = X + X_0 = N$$

$$x = AX = AN - AX_0$$

4. THE X-METHOD WITH AN APPLYING DEVICE

We can design a physical device to apply the above model as X method. We make the above data fuzzy instead of making them distributed. For ex- ample, design a box with two areas A and B. Put some red beans and green beans in it. The numbers are ra , rb and ya , yb respectively. If you are in S, then taking a red bean from A to B, other- wise taking a green bean. Since the information is fuzzy at every step, the privacy of the information is guaranteed. The mathematical model still is

$$\sum_{k=0}^{n} x_k = x + x_0 = N$$

$$x = N - x_0$$

Where X_0 is fuzzy and X_k is unknown to public. The model can be realized in an electronic device

References

[1] Zhang Jingzhong, ”Mathematician’s View” ,China Children Press,2002

[2] Chen Xilu,”Useful Statistics ”,Science Press, 2002

[3] Sun shanze, The basic methods on sensible question survey, Probability Section, Beijing University Lectures 1999