ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research (ASSEHR), volume 106
International Conference on Education Science and Economic Management (ICESEM 2017)

A New Method of Information Fusion and System Optimization in Statistic Inference

Xiao Xiaonan
Xiamen University Tan Kahkee College
Xiamen 363105
xiaoxn@xujc.com

***Abstract*—Sampling design and investigation play a significant role in statistic inference. In order to overcome the defects in traditional method of sampling design and investigation, the article implements multi-index Fuzzy dynamic cluster analysis and offers a unique way of information statistics and packet optimization. In this way the internal information in the sample will be sufficiently explored, thus providing a unique method in statistical analysis and system optimization that will improve the reliability of overall sample grouping and reduce the sampling error on a maximal scale.**

Keywords—statistic inference; information fusion; sampling design and investigation; statistic analysis; system optimization; Fuzzy dynamic cluster analysis; sampling error

I. INTRODUCTION

In the sampling investigation, the multi-index Fuzzy dynamic cluster analysis is introduced into sampling and a way of information statistics to group and optimize is given so that it can overcome the defects many fuzzy and complex factors of sampling being unreasonably grouped in the traditional stratified sampling information processing[1-3]. Combining qualitative sampling analysis with quantitative sampling analysis organically and closely, it offers a new methodology about statistics analysis and optimization for improving the reliability of population sample packet and reducing the sample error furthest.

II. THE DEFECTS AND OPTIMIZATION OF THE TRADITIONAL SAMPLING INVESTIGATION METHODOLOGY

Sampling investigation is not only the method to collect statistical data but the one to estimate and judge the appearance of the population scientifically. In addition, it is of great importance in the statistics analysis and statistics pre-decision study. What's more, it is applied extensively to the modern enterprise management with preferential plan, investment decision, efficiency evaluation and various fields of scientific technology [4]. While there still exists a great many factors in the investigation and improvement on ies and methods [5-6]. In terms of the traditional sampling investigation method, it is to randomly choose part of the entire objects for investigation. Moreover, based upon the acquired data, we could make somewhat reliable estimation and judgment on the quantitative feature of the entire objects so that all of the studied objects can be recognized [7]. However, there are several problems in sampling investigation. For one thing, in the sampling project and the signs of sampling indicators we may have difficulties in taking into account various factors thoroughly which may have an impact on the sampling target. As a result, the research may lose a few helpful statistic information, especially the very complex information such as gray model, fuzzy model, fuzzy and random model, fuzzy and gray model and so on. It may lead to a short of sufficient scientific reliability in the investigation. For another thing, when the interclass unit is adjusted by interclass density, the packet will be not accurate and too sketchy for the reason that it has used the similar method which seems to have made a strong assumption [8-9] .To solve such a statistics methodology problem quickly, the article will apply the multi-index Fuzzy dynamic cluster analysis to conduct the sampling optimization classification study.

III. THE STATISTICS SAMPLING METHOD BASED ON MULTI-INDEX FUZZY DYNAMIC CLUSTER ANALYSIS

A. Clustering principle and method

Because the objective things may contain gray nature and fuzziness in most cases, applying the multi-index Fuzzy dynamic cluster analysis to classify and investigate the sampling makes the grouping more practically.

Supposing $U=\{u_1,u_2,\ldots,u_n\}$ as the unit to be grouped, every sample has m statistic indicators (flag values). xik stands for the number k flag value of the number i sample. What is the result of applying the multi-index Fuzzy dynamic cluster analysis to the above? The key is to make the statistic indicators to be the preferential choice. In other words, statistic indicators should have a clear and real meaning as well as the stronger discrimination, representation and broader meaning.

After choosing the statistic indicator, the multi-index Fuzzy dynamic cluster analysis of the population sampling can be classified as the following three steps:

Firstly, standardize the statistic indicators of each classified sample so that they can be analyzed and compared reasonably. And the formula of the standardized value is:

$$x'_{ik} = \frac{x^0_{ik} - \bar{x}_i}{s_i} \qquad (i=1,2,\cdots n;\ k=1,2,\cdots m) \quad (1)$$

In it, x_{ik}^{0} is the primary data, is the average of the primary data, so

$$\bar{x}_i = \frac{1}{m}\sum_{k=1}^{m} x_{ik}^{0}$$

S_i is the deviation of the primary data, so

$$s_i = \sqrt{\frac{1}{m}\sum_{k=1}^{m}(x_{ik}^{0} - \bar{x}_i)^2}\ .$$

Reducing the standardized value to a closed interval [0, 1], we can calculate the extreme, standardized value in terms of the formula:

$$x_{ik} = \frac{x'_{ik} - \min(x'_{ik})}{\max(x'_{ik}) - \min(x'_{ik})}$$

$$(i=1, 2, \cdots, n;\ k=1, 2, \cdots, m) \qquad (2)$$

Secondly, calculate the statistic which is used to measure the similarity between the two classified samples r_{ij} $(i,j,=1,2\ldots,n)$, and then create the similar relationship based on the population $U=(r_{ij})_{n\times n}$.

r_{ij} can be calculated following the formula:

$$r_{ij} = \frac{\sum_{k=1}^{m}|x_{ik} - \bar{x}_i||x_{jk} - \bar{x}_j|}{\sqrt{\sum_{k=1}^{m}(x_{ik} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{m}(x_{jk} - \bar{x}_j)^2}} \qquad (3)$$

In the formula,

$$\bar{x}_i = \frac{1}{m}\sum_{k=1}^{m} x_{ik}, \bar{x}_j = \frac{1}{m}\sum_{k=1}^{m} x_{jk}$$

There is another formula can be used for calculation according to the actual conditions. For instance:

$$r_{ij} = \frac{\left|\sum_{k=1}^{m} x_{ik}x_{jk}\right|}{\sqrt{(\sum_{k=1}^{m} x_{ik}^2)(\sum_{k=1}^{m} x_{jk}^2)}}; \qquad (4)$$

$$r_{ij} = \frac{\sum_{k=1}^{m}\min(x_{ik}x_{jk})}{\frac{1}{2}\sum_{k=1}^{m}\sqrt{x_{ik}x_{jk}}}; \qquad (5)$$

$$r_{ij} = \begin{cases} 1 \\ 1 - C\sum_{k=1}^{m}|x_{ik} - x_{jk}| & \text{While } i=j \end{cases} \qquad (6)$$

While $i \neq j$

Specifically, C should be chosen properly, and make $0 \le r_{ij} \le 1$ and so on.

Thirdly, cluster. Applying synthesis to transform the Fuzzy similar matrix into the Fuzzy equivalence matrix, enable $\underset{\sim}{R}^{2k} = \underset{\sim}{R}^{k} = \underset{\sim}{R}^{*}$. Then R^* should be clustered and analyzed. Thereby, a realistic classification will be gained according to the arrangement of 0, 1 among R_λ.

Generally, when the number of samples tends to be quite large, we can build the similar Fuzzy relationship $\underset{\sim}{R}$ among the samples and synthesize them for several times n terms of the characteristics of the samples to be clustered. Then, we can transform into the Fuzzy equivalent matrix $\underset{\sim}{R}$.What's more, all this can be completed by the electronic computer. It is not only accurate but also fast to adopting the electronic computer to the population sample with the multi-index Fuzzy dynamic cluster classification. For instance, if there are 100 samples and every sample has 10 indicators, it just needs 3 minutes or so from starting calculation to outputting the Fuzzy equivalent matrix. The accurate and instant level is what the traditional stratified sample cannot reach.

B. *The test of the classification result*

Supposing n is the number of the population unit which has been grouped and divided into t types. In order to test the rationality of the classification result, the statistic can be selected following the distribution $\eta\ (m, n\text{-}t, t\text{-}1)$. And E is the within-group scatter matrix. In addition, its element follows the formula

$$\eta = \frac{|E|}{|W|} \qquad (7)$$

$$e_{ij} = \sum_{p=1}^{t}\sum_{q=1}^{n(p)}(x_{qi}(p) - \bar{x}_i(p))(x_{qj}(p) - \bar{x}_j(p)) \qquad (8)$$

In the formula, $n\ (p)$ is the number of the population units of the Group p. $x_{qi}\ (p)$ is the ith flag value in the qth unit of the Group p, $\bar{x}_i(p)$ is the ith average of the Group p, $x_{qi}\ (p)$, $\bar{x}_j\ (p)$ follows the rules as well. What's more, i, j=1, 2,…, m while W is the total deviation matrix. And its element follows the formula:

$$W_{ij} = \sum_{q=1}^{n}(x_{qi} - \bar{x}_i)(x_{qj} - \bar{x}_j) \qquad (i,j=1,2,\cdots,m) \qquad (9)$$

As for the given significance level a, if $\eta < \eta_a\ (m, n\text{-}t, t\text{-}1)$, then it shows that there exists significant differences between each group indexes. Therefore, the groups are classified reasonably.

C. *To determine the number of the samples*

After the groups are determined, the interclass variance is fixed and the mean-square deviation of the jth index of the pth group is:

$$S_j(p) = \left(\frac{1}{n(p)-1}\sum_{q=1}^{n(p)}(x_{qj}(p) - \bar{x}_j(p))^2\right)^{\frac{1}{2}}$$

$$(j=1, 2, \cdots, m) \qquad (10)$$

Supposing $n_j(p)$ is the number of samples extracted from the pth group with the non-repeated sampling according to the jth index, the variance of the sample mean in the jth index of the pth group is

$$\left(\sigma_j(p)\right)^2 = \frac{1}{n_j(p)}\left(S_j(p)\right)^2\left(1-\frac{n_j(p)}{n(p)}\right) \quad (11)$$

Therefore, when it extracted according to the jth index, the mean error is

$$\mu_{\bar{x}_i} = \left\{\sum_{p=1}^{t}\frac{(n(p))^2}{n^2}\cdot\frac{\left(S_j(p)\right)^2}{n(p)}\cdot\left(1-\frac{n_j(p)}{n(p)}\right)\right\}^{\frac{1}{2}} \quad (12)$$

In the unequal proportional sampling if the total of the sample is fixed, in order to minimize the error of the sampling, then

$$n_j(p) = \frac{n(p)S_j(p)}{\sum_{p=1}^{t} n(p)S_j(p)} n_j \quad (13)$$

In it $n_j = \sum_{p=1}^{t} n_j(p)$

If the limited error of the jth index is Δ_j, then

$$\Delta_j = u\mu_{\bar{x}_j} \quad (14)$$

When the guarantee is 1-a, u is the critical threshold of the normal distribution. Substituting formula (12) and (13) into formula (14), we can get the samples when they are extracted according to the jth index after the population is divided into t groups.

$$n_j = \frac{u^2\left(\sum_{p=1}^{t} n(p)S_j(p)\right)^2}{n^2\Delta_j^2 + u^2\sum_{p=1}^{t} n(p)\left(S_j(p)\right)^2}$$

$$(j=1, 2,\cdots, m) \quad (15)$$

In order to meet the precision requirement of the m indexes, the t groups samples can be

$$n(t) = \max_{1\le j\le m}\{n_j\}$$

In fact, combining our experience we may define a scope for the groups. That is $t_1 \le t \le t_2$. Then we can select the smallest samples from all the groups to be the final result.

$$\min_{t_1\le t\le t_2}\{n(t)\}$$

D. The example of the sampling design

Through the investigation of the production and sales of a number of Xiamen enterprises, we have determined 1456 enterprises to be the population for the sampling investigation. Next, we have selected the 10 catergories for the features of the groups which are highly related to the production and sales: the total industrial output value, the net industrial output value, the fixed asset value, the normal sales profits, the maximum profitability in busy season, the off-season maximum loss, the profit and tax value, the input of education and scientific research, the input of operating management, the number of staff and so on.

Because using 1456 samples directly for the Fuzzy dynamic cluster analysis will cost much energy and time, we extract 210 units from the population randomly to cluster and set up the Fuzzy similar matrix R^* between samples according to their characteristics. All these can be completed by electronic computers. Besides 1456 population units can be classified into 8 groups by the Fuzzy dynamic cluster. Consequently, under the condition that the allowable error is 5% of the index average and the guarantee 95%, we would figure out the number of the required samples is 264. Then, we can draw 264 samples randomly from 1456 population units without repetition so that we can make a precise statistic inference for the production and sales level according to the sampling information [10-12].

IV. CONCLUSION AND THE PROSPECT

The article uses the method of the statistic information sampling investigation with an emphasis of the multi-index Fuzzy dynamic cluster analysis in order to improve and develop the sampling investigation method. There are several advantages to group the population samples by the multi-index Fuzzy cluster analysis. First, the internal information will be discovered substantially. Second, many fuzzy and complex factors of the samples can be considered thoroughly. Third, it will overcome the defects of the approach of the traditional stratification sampling whose information is unilateral. Fourth, it may combine the qualitative sampling analysis and quantitative sampling analysis closely. Last but not least, it offers a new way of statistics analysis for optimization in order to improving the reliability of population sample packet and minimizing the sampling errors.

REFERENCES

[1] Xu Baolu, *A Theory of Sampling* [M]. Beijing: Beijing University Press, 1982(In Chinese).

[2] Sun Shanze, *Sampling Investigation* [M]. Beijing: Beijing University Press, 2004(In Chinese).

[3] Xiao Xiaonan, *Modern Information Decision Methods* [M]. Beijing: Beijing University Press, 2006(In Chinese).

[4] Du Zifang, *Sampling Technique and Application* [M]. Beijing : Tsinghua University Press, 2005(In Chinese).

[5] W.G. Kirk, *Sampling Techniques* [M]. Beijing: Chinese Statistics Press, 1985(In Chinese).

[6] Zhao Junkang, *The Theory and Method of Sampling Design in Statistics Investigation* [M]. Beijing: Chinese Statistics Press, 2002(In Chinese).

[7] Chen Xingxin and Cao Min, *Statistics Calculation Method* [M]. Beijing: Beijing University Press, 1989(In Chinese).

[8] Kalbfleisch J G. Probability and Statistical Ieference[M]. New York: Springer-Verlag,1985.

[9] Morris C N, Polph J E. Introduction to Data Analysis and Statitical Inference[M]. New Jersey: Prentice-Hall, 1981.

[10] Du Dong, *Statistics Information System* [M]. Beijing: Chinese Statistics Press, 2006(In Chinese).

[11] Wang Yufang and Xiao Shitang, *The Planning and Process of Statistics Sampling Investigation* [M]. Beijing: Chinese Economy Press, 2005(In Chinese).

[12] Xiao Xiaonan. The Minimax Estimation of Corresponding Risk of Statistical Model. Journal of Xiamen University (Natural Science), 2008,47(5) : 641−643.