Using Cluster Analysis in Stratified Multi-Variated Sampling

Ruan Dacheng
Room 1408,No.48,Weihai Rd.,Shanghai,China
ruandc@email.stats-sh.gov.cn

Abstract: Stratified multivariable sampling and cluster analysis is apparently different branch of mathematical statistic. But through study, we find that the Ward' s method of cluster analysis can be used todealwithstratifiedsamplingsurvey.WecanfirstlyuseWard' smethodtodoclusteranalysisonmultivariate population and get a certain quantity of clusters, then let these clusters be the strata of stratified sampling. As a result, the precision of multivariate sampling survey can be improved, which is the aim of this paper.

1.The stratified sampling survey method

In the stratified sampling, the sample population with N units is firstly divided into several subpopulations with $N_1, N_2 \ldots N_L$ units respectively. Those subpopulations have no repetition and can be combined to be the sample population, i.e. $N_1+N_2+\ldots+_{NL}=N$. The subpopulations are called strata. Then, samples are taken from each stratum independently with $n_1, n_2, \ldots, n_L$ showing theirsizes. If simple random samples are taken from each stratum, the method is then called stratified random sampling.

Stratified random sampling is a well-used survey method. The purposes that we choose this method are omitted:

The fundamental methods of stratified random sampling are the following(omitted):

Then, how to level strata to make it more appropriate? In common there are two methods. One is to level the strata according to some features of the population, the second method is to take the variable y needed to estimate in the sampling survey as the stratified variable, arrange the population in order according to this variable, and level the population into several strata, in order to no repetition between each stratum. As to this variable, the variance inside the stratum will be smaller than that of the total. Obviously, the second method is more advanced than the first one.

However, thinking of the cost of sampling surrey and other causes, we seldom use only one target in one sampling surrey, instead, we try to investigate more targets. Thus we have to face one problem--how to control several targets (i.e. variables) at the same time ?

2.Cluster Analysis

Cluster analysis is a method to be used in "classification" cluster in mathematical statistics. It can also be viewed as a divided branch in multivariable statistical analysis. The basic method of it is: (In this study we cluster only according to distance) to define the distance in m-dimension space, and to classify the dots that are close to each other into one cluster and those far toeach other are classified to different cluster.

The definitions of the distance are different according to the different types of variables. We mainly deal with the scale of the distance, which means the target is remarked by continuous quantity. Let X_{ij} bethejthtargetoftheithobservation,andd_{ij}bethedistancebetweentheithobservationandthejthobservation. The distance should generally meet the following four terms(omitted):

In Cluster Analysis, the distance in common use is the following(omitted):

In Cluster analysis, Hierachical Clustering Methods is well used at present. Its main idea is: take the n observations as each cluster respectively, and define the distance between observations and clusters. Initially, because each observation is a cluster, the distance of observations is as same as that of clusters. Choose the couple of observations that have the shortest distance to unite that have the shortest distance to unite into a new cluster. Calculate the distance between the new cluster and the other, then combine the two nearest clusters. So every time, the amount of clusters will reduced one until all the clusters are

combined into one cluster.

According to the different definition of the distance between clusters, Hierachical Clustering Methods is mainly classified as following(omitted):

Suppose that n observations are classified k clusters, G_1, G_2 ..G_k. Let x_{it} be the I-th observation of G_t (notice :x_{it} is a vector in m-dimension space), n_t be the number of observations of G_t, and x_t be the centroid of G_t, then the sum of square deviation of all observations in G_t cluster is :S_t=(omitted)
and the sum of square deviation of all k clusters is: S=(omitted)
when the number of clusters k is definite ,we want to choose the right classificationthat makes s minimum. As for n observations are classified into k clusters, the quantity of all possible classifications is (the demonstration is omitted)R(n,k)=(omitted)
Usually it is impossible to choose minimum s through comparing so many classifications. So we have to work out some calculating rules to find out a local optimal solution. Ward' s method is just such a method to looking for local solutions. Ward' s idea is: At first let each of n observations be a cluster, then reduce one cluster each time. And the sum of square deviation will increase each time when reduce a cluster. Let the two clusters that increase S least be combined into one cluster, till all observations are in one cluster. Then, what is the distance in the sum of square deviation? In fact, in this method the sum of square deviation increasing when two clusters are combined is considered as square distance. So, Ward' s method can be combined as unified with other hierarchical clustering method.

3.The application of cluster analysis in the stratified multivariate sampling

We can cluster the sample population by the sum of square deviation, then let the clusters be the strata of the stratified multivariate sampling. Therefore we can effectively control the precision of multivariate sample survey.

During the cluster analysis for the stratified multivariate sampling, at first, we should standardize variables in order to get rid of the affection of different prickle. Considering that the size of the target population is big, it will cost a lot of time and memory if we use the Ward' s method directly, so we can firstly cluster the target population by k-cluster method. We look on the result as the initial cluster result. Second, we cluster the result (the initial cluster result) by the Ward' s method. At last, the result is the strata needed by stratified multivariate sampling.

4. Brief Report on the result of the practical application

From 1997 shanghai has launched a pilot project on the sample survey of those industrial enterprises under certain scale. In that survey, we adopt the above cluster analysis method in stratifying and adopt 95%confidenceandthe10%minimumrelativeerror.Theaccuracyofthecomputedresultsisverysatisfying. With view to the distribution of those industrial enterprises in shanghai, sample population' s status quo and the feathers of administrative system, we divide Shanghai industrial enterprises under certain scale into two subpopulations, respectively named subpopulation É and subpopulation Ð. Separately, we use the above-mentioned methods to sample the two subpopulations, investigating several targets, such as the year-end staff number, industrial gross product, product sales revenue, etc. The actual results of the three years' sampling investigation from 1997 to 1999 show that stratifying may increase the accuracy of the estimated values of population targets. The sampling precision of most qualifications is under 5%.