Studies in Multivariate Stratification: Similarity Analysis vs Friedman-Rubin

David J. Fitch, dfitch@uvg.edu.gt, and Roberto Molina Cruz, rmolina@uvg.edu.gt,
Universidad del Valle de Guatemala

1. Introduction

McQuitty (1957) observed that if classes of people, PSUs in sampling terms, were formed where PSUs within strata were identical with regard to stratification variables, the number of such variables that could be used would typically be much smaller than the number available with relevant information. He proposed Similarity Analysis. This was programmed for the ILLIAC (Fitch,1958). Starting with a matrix of similarity scores, the two most similar PSUs or PSU group are combined at each iteration to form a developing stratum. At the 1969 ISI meetings, Beale (1969) presented an alternative stratification procedure. Starting with random strata, each PSU of each stratum is tried in each stratum and left in that stratum that gives the largest between sums of squares, iterating until the gain is small. Census is using a similar method in their CPS design (Friedman and Rubin, 1967). We will refer to this as the F-R method. We have programmed a variant of each method for large, fast PCs, and have planned a series of studies on the utility of each and with variations and combinations. For the work here reported we are using data for the 220 variables from the 1994 County and City Data Book (www.census.gov).

The programming of the two stratification methods was for us a difficult and lengthy process. Some 2,500 lines of Fortran programming were written. Happily we now have a result from a first comparison, but there are unanswered questions that await further work. We will need to try to better understand and analyze problems and possibilities. Hopefully you can help us with your questions and comments. We have done some things that seem to us non-standard, but as far as we can see are theoretically correct. We will give details on these in the hope of clarifying our thinking and in an effort to get help from you. We will begin with some details of the data used in the comparison. Next we will describe the programs. Following this we will present our first results. Finally we will give some hopes and plans for further thinking and work on multivariate stratification, or what we will argue might productively be thought of as nonlinear stratification.

2. Data

The data used in the comparison were both realistic, and artificial. They were from the 1994 County and City Data Book CD published by the US Census Bureau for the 3141 counties and county like divisions of the US. This CD contains, with a small amount of missing data, observations for 220 variables for each county. These county data are in 20 files. We used Epi Info 2000 (www.cdc.gov) to prepare files which could be read by Fortran. Programs were written to input these 20 files and combine them into a 3141 by 220 data matrix. Next simple imputation was done.

Perhaps this is the place to note a basic philosophy in our work. Our goal is not to make better estimates for real variables for real populations. Our goal is to seek efficient stratification methods. To have any hopes of making a beginning we need to give up any hopes of making real estimates where one would have to select PSUs that differed in size and then sample within selected PSUs. And in the real world, of say the US Census Bureau, there are a multitude of considerations that must be taken into account. Making state estimates and selecting PSUs so as to facilitate the work of interviewers would be two such considerations

.

Some of the 220 variables were highly skewed such as base populations - there being some very large counties. It seemed reasonable to reduce some of this skewness but in the interest of getting a data set on which to develop our programs, and in light of the philosophy expressed above, this was done in only a crude fashion. Next the variables were standardized to a mean of 0.00 and a variance of 1.00. We ask you to think with us as to the implications of this handling of our data. The variables in the CCDB are grouped as to subject matter. So as to get, in the first 20, a set of stratification variables which would likely account for a good range of the total variance, the order of the 220 variables was randomized. We have now described our data set

2. Our Programming

Programs were written in the language of Essential Lahey Fortran 90, ELF90 (www.lahey.com). This has a lot to be recommended including its price - $149 for university people. However we have found that it has some errors. Lahey, although they still sell the program, is not correcting errors that are now found, so it would probably better to purchase their current Fortran. The minimum price is $249. We think that we have worked around any errors, but it remains a source of concern. Perhaps we will have bought a new compiler and have compiled with it in advance of the meetings.

3.1 The F-R method

This method begins with a set of strata, say 100 as in the present case. These could be where PSUs were

randomly assigned to initial strata, which is what we did and what the CPS of the Census design, as we understand, does. The method iterates through these 100 strata, and within each stratum, takes in turn each PSU of the stratum and tentatively puts it in each of the 100, computing the between sums of squares sum with the PSU in each stratum. Alternatively the within sums of squares sum with the PSU in each stratum, might be computed. As the within sums of squares plus the between sums of squares is a constant – the total sums of squares – one can either aim to maximize betweens or minimize withins. After the PSU is tried out in each stratum it is put into, or left in, that stratum which, if one works with betweens, gives a maximum. As we were working with 20 stratification variables we had for each trial switch 20 between sums of squares and made out switching decision on the basis of the sums of 20 betweens. In programming this F-R method we first, as a check on our programming, computed in **pfr1** both between and within sums of squares checking to see that we got the correct total in each case. Then we reprogrammed, in **pfr2,** basing our switching only on the maximum between sums of squares, checking the resulting stratification against that of **pfr1**. We found initially some differences. These disappeared when, for testing, we worked with double precision with each. **pfr2** is some 40 times as fast and takes in fact very little time to do 5 iterations which is all that one needs to get very close to a local maximum. Leaving the computer running on a weekend could give thousands of solutions with different random starts. The initial program was modified to output the data files that are then input by **ps2,** our stratification program using the similarity method.

3.2 The Similarity Method

First, what became **ps2,** was first written to do only the similarity analysis. There are a number of test options built into the program but for our work here we used the option for doing a regular analysis. A program to be described later created the data sets for input in what is called list-directed formatting. After asking to choose an option, the number of PSUs, the number of stratification variables, and the number of strata to be formed, the program asks whether the user wants to use covariances or a distance measure as the index of similarity – the square root of the sum of differences squared. We use the later although we want to think about possibilities for the use of covariances. We will first describe the original program. It computes a triangular matrix of similarity indices. With a large number of PSUs this takes a lot of RAM memory and was a motivation for getting a computer with a large amount of such memory. But memory is cheap. We paid like $150 for a gigabyte. Then begins a series of iterations. The largest similarity, i.e., the smallest distance, is found and the two counties or county groups are combined to form a new grouping, i.e., a developing stratum (DS), and the mean similarity between this DS and all other counties or DSs is computed and the next iteration initiated. The program was written to facilitate a sampling design that uses two PSUs per stratum, and small weight variance. Although not relevant here, as we make all PSUs the same size, the program uses PSU size figures, i.e., number of dwelling units (DUs), and in the stratification aims at forming strata which are approximately equal in total number of DUs. The program forces, when PSUs are selected with probability proportional to size (pps), the selection of at least two PSUs and not more than three. Three are selected 20% of the time. In a two PSUs pps selection option, two PSUs are selected with accompanying weights. In the present work we selected one PSU per stratum and used the stratum weight given by the program.

Built into this first phase of the program - that iterates through the triangular matrix - are some controls on sample size. These and later such controls may give difficulties not presently understood, but as noted our proof of method superiority is in the accuracy of estimation from the stratification produced by the method, not in the theoretical elegance of the method. At some point the first phase is terminated. Having specified 100 strata, as in the present case, the largest 100 DSs are located and are the columns of a rectangular matrix that is now formed and which will be used to complete the stratification process. The rows are the single PSUs that have not entered into any of the column strata. Each element of this matrix is the mean similarity, weighted for PSU size, between the row PSU and each PSU of the column strata. Now row PSUs are added to their most similar column stratum, with stratum size controls. At the end of this process these size controls are relaxed, the only control being that no stratum can grow in size beyond the point where the pps selection might select more than three PSUs per stratum. It would have been much easier to have programmed similarity analysis without any size control but we wanted, do want, to try to introduce the two PSUs per stratum design into places such as Guatemala where it is largely unknown. And the fact that our programming of what we call the F-R method has no size stratum controls means that our comparisons are not as comparable as they might have been. But we hope that our two PSUs per stratum work will be, as we say in Guatemala, “Vale la pena” – “Worth the worry”.

We found in using this program that something was wrong. With variables standardized to a variance of 1.00, the mean of a sample of size 100 would be expected to be 1.00/100 = .01 but we got like .015, so there obviously was an error. Although we worked right up to the time the paper was to be presented, it was too late to find the error. Two aspects of the original programming were complicated and likely we will find

the error in one of these two parts – working with a triangul ar matrix, and procedures to control for stratum size. Both were removed and the program redone. This meant that with this programming, stratum sizes, as with the F-R method, varied. A new feature was added in this, not included in the F-R program. With stratum size variation, and selection of PSUs to be done systematically, this meant the no selection would be made from some small strata. The strata were reordered, with the second stratum the one remaining most similar to the first, the third most similar to the second, etc. This gave a somewhat unfair advantage to the similarity analysis method.

3. Results

Estimated variances for the estimated means for the remaining 200 variables, i.e., 220- the 20 variables used to stratify, were made as follows, with the procedure being the same with both methods. Counties, 41, were randomly eliminated, leaving 3100. Systematically 31 samples of 100 counties were formed, and sets of 200 means were computed for each sample. The equation $v(\bar{y}) = \sum_{i=1}^{31} (y_i - \bar{y})^2 \big/ 31$, where y_i is a mean based on 100 observations and $\bar{y}$ the overall mean, to estimate the gain over .01 with each stratification method. With the similarity analysis method, possibly still with errors in the program, the estimate was .009 and with the F-R method the estimate was .005. So our present, but tentative, conclusion is that the F-R method is the considerably better method.

4. Discussion Points

4.1 A problem in the similarity analysis program
Thinking it might not be clear that a sum of one observation per similarity stratum, weighted by the number of PSUs per stratum - all PSUs being of the same size - we undertook the following demonstration. We generated a data set with 1,000 observations, 10 each, - 1.,2., …, 100, and randomized these data. The idea was to have shown that a random selection, one PSU per stratum, would give a mean of exactly 50.5. We found that the similarity program stratification, although very close to the perfect expected, was not perfect. We will try to find the problem. We may find that, in order to meet other goals, it is not be possible to get the stratification expected here.

4.2 One PSU per stratum
Although we are less familiar with the stratification literature than we should, it is our impression that sampling statisticians are increasingly using one PSU per stratum. It would seem to be useful to compare variance estimates from say 50 strata using ordinary variance estimation techniques with those from 100 strata. We have, as we are working with population data, variance estimates based on 31 estimates, each based on 100 strata. We hope to develop programming that will pair similar strata, and compute variance estimates with 50 strata which we will then compared with those from 100 strata. Perhaps such analyses will help better understand the variance situation when one selects only one PSU per stratum.

4.3 Non-linear stratification
As surveys are typically undertaken to estimate a range of variables we want to stratify so that the sampling units are similar on a combination of variables. Let us consider the Meehl paradox (Meehl, 1950). He said let's assume two test items. Normal people respond 11 or 00 to these items while schizophrenics respond 10 or 01. Here, where we score using $T = 1 - x_1 - x_2 + x_1 x_2$, we get a score of 1 for normals and a 0 for schizophrenics. Without the cross product term, i.e., with a linear equation there would be no way to form the two strata of people using these test items. It is our belief that thinking in terms of such nonlinear stratification will lead to important developments, although how these developments will look is not at all clear to us.

4.4 The F-R method controlling for stratum sizes
The original similarity analysis method controlled for stratum size. If we find the error in this method and then get an estimate based on strata of approximately equal size, we might explore modifications of the F-R method to give strata of approximately equal size. Approximately equal size strata would allow the selection of two PSUs per strata with approximately equal weights, and thus would encourage the use of the balanced half sample method, something we would like to see happen in Guatemala.

4.5 The F-R method starting with different initial strata
It seems to us that there must be better starts than random assignments. One could, e.g., start a F-R stratification with a similarity stratification, or the results of using some other clustering method.

4.6 Two or more stages of stratification
It seems to us that there would be more to gain from stratifying within PSUs than stratifying PSUs such as counties in the US. For example one might stratify using Census tracks within

selected counties, and then possibly by block within selected tracks. We are currently undertaking such work based on the now available data from the 2000 US census.. But as pointed out by (personal communication) some Census tracks will over time undergo considerable size changes. Such could mean considerable weight variance which would reduce efficiency.

4.7 The complexities of the real world here not considered

Our comparison of our two methods was greatly simplified by the assumptions that all PSUs were of the same size and that all stratification variables and the variables to be estimated had a mean of 0.00 and a variance of 1.00. As we think about how further work in stratification might go, it seems reasonable to continue to undertake studies with such simplifications, leaving to later the adding of complexities of the real world to later. One then might use those stratification methods found useful in studies using such simplified data.

5. References

Beale, E.M.L. (1969), Euclidian Cluster Analysis. Contributed paper to the 37th session of the International Statistical Institute.

Fitch, D. J. (1958), Predicting voting behavior of senators of the 83rd congress: A comparison of factor analysis and similarity analysis. Unpublished doctorial dissertation, University of Illinois, Urbana.

Friedman, H. P. and Rubin, J. (1967) On some invariant criteria for grouping data. Journal of the American Statistical Association, 62, 1159-1178.

McQuitty, L. L. (1957), Isolating predictor patterns associated with major criterion patterns. Educational and Psychological Measurement, 17, 3-42.

Meehl, P. E. (1950). Configural scoring. Journal of Consulting Psychology, 14, 165-171.