NONLINEAR STRATIFICATION

David J. Fitch
Instituo de Nutrición
de Centro América y Panamá,
dfitch@incap.org.gt

In the 1950's there was considerable interest within psychology in the classification of people into classes who behaved similarly - say who responded similarly to a set of items of a psychological test. In sampling we would call these classes strata. Meehl (1950) had published his famous paradox. He said, let us imagine that we have a test with two items. Normal people respond 11 (yes-yes) or 00 (no-no) to these items, while schizophrenics respond either 10 or 01. Here he said we had two items that separated normals and schizophrenics perfectly but how might we score these two items, giving one score for normals and another score for schizophrenics? Horst, slightly chiding his fellow psychologists for their lack of mathematics, published (1954) the solution. The answer to Meehl was to use a nonlinear equation, one with, in this simple case, a single cross product term. The equation $y = 1 + 3x_1 + 2x_2 - 4x_1x_2$ gives, for the two normal types 1 and 2, and for the two schizophrenic types 3 and 4. So nonlinear stratification would be expected to give strata where measures within would be similar and between different, as compared with stratification based on a linear function of the classification variables.

Kish and Anderson (1978) in their important stratification paper give related conclusions. They show that under reasonable conditions where both a x_1 and a x_2 variable are each divided into 6 ranges and 36 strata are formed, improved estimation follows, as opposed to forming 36 strata based on a linear function of the two variables. A problem in practice would seem to be how one would pick, and divide, the stratification variables. Obviously the number of stratification variables in this case would be limited. If we were to use 4 variables each divided into 6 ranges we would have $6^4 = 1296$ strata and such numbers would usually not be practical. The situation is analogous to problem on which those in psychometrics, such as McQuitty (1957) worked in the 1950s. He called grouping people who behaved identically on a set of items, pattern analysis, and grouping people who behaved similarly, similarity analysis.

So how might we form strata in a practical way that responds to the fact that we live in a nonlinear world - that people within PSUs that in many ways are quite different can in some ways respond similarly, and that people within PSUs that are quite similar in some ways can respond differently? What indices would usefully show similarity, and how might we use such to group PSUs into strata. I would like to suggest a variant of McQuitty=s similarity analysis, sed to classify Senators (Fitch, 1958) programmed for the ILLIAC. This 1950's computer with 1024 words and some auxiliary storage was strained nearly to its limit classifying 88 Senators. Programming, done more recently on a computer with 384 MBs of RAM, can form a specified number of strata, selecting 2 PSUs pps per stratum, starting with 14,000 PSUs. We have here in Guatemala something less than this number of census enumeration districts, which led me to acquire this RAM.

To introduce the method let us return to our opening situation with 11, 00, 10, and 01 people. But now let us think of them as PSU to be grouped into strata. If we have two of each type, 8 PSUs in all, and computed a triangular matrix of correlations between these eight we

PSU		1	2	3	4	5	6	7	8
		11	01	10	00	11	01	10	00
1	11								
2	01	0.							
3	10	0.	-1.						
4	00	-1.	0.	0.					
5	11	1.	0.	0.	-1.				
6	01	0.	1.	-1.	0.	0.			
7	10	0.	-1.	1.	0.	0.	-1.		
8	00	01.	0.	0.	1.	-1.	0.	0.	

would get the results as shown in Table 1.

In this little example PSU 1 and 5 would form the first stratum, 2 and 6 the second, 3 and 7 the third and 4 and 8 the fourth.

Now some details on the first of the two versions of the program. It was written in Essential Lahey Fortran 90, ELF90. The user would put the program and the data into a directory. Data are 1) an N by M matrix of PSU measures, e.g., from a country=s census - with no practical limitation on the size of M, 2) a vector which is the judged importance for the purposes of the intended survey of each of the stratification variables, and 3) a vector which is the size of

each PSU. With the user supplying the parameters N, M, and the number of strata NS to be formed, a triangular matrix of covariances between PSUs is computed using the importance weights supplied. In the first part of the stratification procedure at each iteration those two PSUs or PSU groups with the largest covariance are joined and the mean of their covariances with each of the other PSUs or PSU groups replaces the covariances of the first joining PSU or PSU group and zeros are entered for the second. In this first part there is some control on the size of each developing stratum. The second part of the stratification procedure takes the largest NS groupings and adds, one by one, the unassigned PSUs, attempting to place them in the stratum with which they are most similar. Again there is some size control so that all strata are of

approximately equal size. The program outputs two sets of selections. Let us say size figures are number of dwelling units (DUs) per PSU. Both selections are pps. In the first, 2.2 PSUs are selected per stratum, allowing equal DU initial weights, assuming the same number of DUs are selected from each selected PSU. The second selection gives 2 PSUs per stratum meaning that either DU weights would vary somewhat from stratum to stratum, or the number of DUs would need to vary if weights were to be the same in the different strata.

A practical limitation on the number of strata in a large developed country such as the US is the cost of PSU offices. This limitation does not hold in a small developing country such as Guatemala where all survey work is typically out of the capitol. Let us assume that a national survey decided to interview women in 6,000 DUs. And let=s say, as opposed to much work in developing countries, that a reasonable efficient number of DUs per PSU were to be selected - say 10 although 5 might be better. So with 10 this would be 600 PSUs or, with 2 PSUs per stratum, 300 strata. Some day with research in developing countries available, and used, we might see surveys with Atakes@ of 5 DUs and 500 strata.

For purposes of testing, an option was programmed where a degree of correlation was built into successive groups of 50 PSUs, and PSU size figures were generated with a mean of 100 DUs, and a standard deviation of 10. The program gave reasonable results with these data. Next data (correlations) were generated based on two, three value, classification variables, -1., 0., and +1., giving 3^2 =9 possibilities. One might think of these values being below average, about average, and above average. Here the stratification was less satisfactory. So in a second version of the program a distance measure was used as the index of similarity between PSUs. The index d_{ij} is 1.- ((sum of the differences squared between PSU i and j on all of the measures, standardized and importance weighted)/ the largest sum of the differences squared between PSU i and each PSU. In this second version, using the same PSU sizes as noted above, and with equal weighting of the classification variables, assignment to the nine strata was perfect.

As in all, especially new and complicated programming there is the possibility of errors. The fact that this programming, except for the added distance option, was done 2-3 years ago, laid aside, and then now found upon picking it up again, to be without discoverable error, gives encouragement. The real test, and the one I seek, is to try the program out with real data. If one had, say PSU means on 100-200 variables from a country' s census for all of the country' s census enumeration districts, could have access to a sample of census forms from each district, selected some variables to hold out, stratified on the remainder, sampled 2 PSUs from each strata, obtain a sample of forms from each sampled PSU, and estimated population means from such sampling, comparing such estimates with population figures - well I would be very happy to see such done. Such is not likely to happen soon in Guatemala. Our National Institute of Statistics is not into such things. They don' t have any statisticians. I would be pleased to send the programming to anyone interested.

Some Final Thoughts

1. It might be the case that a country had data for each of its census enumeration district (PSUs) on a large number of variables, say 100. These 100 could be used to compute covariances between their PSUs. One might, citing the small increase in cross-validated multiple r' s beyond the use of a few variables conclude that there would be little or no gain from using such a large number of classifying variables. Replying to this, let' s take the case of two sets

of PSUs that in most ways were very similar, but in a few ways were quite different. We might well like to see these two sets in different strata. Perhaps the method here presented would give us what was wanted.

2. A possibility related to the multiple frame approach (Bankier, 1986) would be to use a program such as here described to form a few strata, and then to follow this with a second stage stratification in which each first stage stratum was stratified. Perhaps those variables important in the first stage would be given a lesser importance in this second stage.
3. As noted, controls were placed on the sizes of the strata so that there is not much variance in the stratum sizes. This was done to keep small the weight variance and hence the variance of the survey estimates. However more homogeneous strata would be possible if such controls were relaxed.
4. The controls on stratum sizes programmed were based on intuition and seem to work with the PSU size variance used but would not likely work well with different size variances. It would be good to give attention to this.
5. I see a possibility for more elaborate indices The covariance matrix could be factored. The large size would suggest using the square root method (Fitch, 1958). Factor scores would be computed for each PSU, the distance in terms of these scores between PSUs computed, and stratification done using these distance scores.
6. And finally let me mention a stratification possibility that has intrigued me. One might order PSUss of each stratum on some index of similarity from most to least and then pair the two most dissimilar, then the next most dissimilar, etc. From the $N_h / 2$ PSUs so formed two would be selected.

REFERENCES

Bankier, Michael D. (1986), "Estimators Based on Several Stratified Samples With Applications to Multiple Frame Surveys," *Journal of the American Statistical Association*, 81, 1074-1079.

Fitch, David J. (1958), *Predicting Voting Behavior of Senators of the 83d Congress: A Comparison of Factor Analysis and Similarity Analysis.* Unpublished Ph.D. Thesis, University of Illinois, Champaign.

Horst, Paul. (1954), "Pattern Analysis and Configural Scoring," *Journal of Clinical Psychology,* 1, 3-11

Kish, Leslie, and Anderson, Dallas W. (1978), "Multivariate and Multipurpose Stratification," *Journal of the American Statistical Association,* 78, 24-34.

McQuitty, L.L. (1957), "Isolating Predictor Patterns Associated With Major Criterion Patterns," *Educational and Psychological Measurement,* 17, 3-42.

Meehl, Paul E. (1950), "Configural Scoring," *Journal of Consulting Psychology,* 14, 165-171.