Getting the Most from Demographics:

Things to Consider for Powerful Market Analysis

Charles J. Schwartz

Principal, Intelligent Analytical Services

Demographic analysis has become a fact of life in market research. By linking sales, surveys, and local markets it provides a powerful tool for market planning, segmentation, and target marketing. Readily available in user-friendly PC databases and predigested into lifestyle clusters, it is a quick and relatively painless way to provide quantitative prescriptions for marketing programs.

As is true in almost any kind of research, there are many ways to do demographic analysis. Each has different degrees of difficulty and often different underlying assumptions. These differences will influence both the cost and effectiveness of any demographic based project. This article will look at some of these issues as they relate to demographic market segmentation and penetration analysis.

Population at Risk

In demography, the first step of almost any analysis is defining the population at risk of the event under study. A man cannot give birth, for example, and a 40 year old cannot die of Sudden Infant Death Syndrome. Consequently, comparative fertility studies base their birth rates on the number of women of childbearing age; studies of infant deaths base their rates on the number of persons under 1 year of age.

In market analysis, as much as in formal demography, population at risk is a crucial concept. Market potential must be defined in terms of the population "at risk" of buying the product. This can be in demographic terms - Medicare insurance must go to people over age 65; luxury cars for the most part are sold to those with sufficient incomes to pay for them. It can also be in terms of factors related to previous marketing efforts. A bank's customers are by and large confined to those living within the trade areas of

its branches and responses to direct mail campaigns can come only from those who received the mailings.

While population at risk sounds basic, it is often ignored. For example, the canned lifestyle cluster analyses available from most vendors will compare the percentage of sales by cluster to the percentage of the population in each cluster. The higher the ratio of the sales percentage to the population percentage, the more favorable is the cluster. If the product under analysis is relevant only for a particular group - those with high incomes or those over age 65 - then all the cluster analysis will succeed in doing is identifying clusters with high percentages of their populations in the relevant age group or income category. It would be better to use age or income directly and then think about cluster analysis.

If you have prior knowledge of the demographics of your market, use it. Such knowledge can be structural (only those over 65 are eligible) or it can be based on clear results from primary or syndicated research (90% of respondents who would buy the product have incomes over $50,000). Instead of using the entire population as your basis for comparison, use the population over age 65 or the percentage of those with incomes over $50,000. You may find that your Medicare supplemental insurance sells to retirees who live in non-retirement communities or that your luxury car sells best to people whose incomes are high in comparison to the income of their neighbors. Combining this information with the size of the population at risk will result in more accurate market potential measures.

In addition to demographics, location limits population at risk for any product or service that relies on a network of retail outlets or branches. In many cases this factor can be controlled by limiting the analysis to the trade areas of the outlets, but often such controls are inadequate. For example, a

company may have targeted particular demographic groups in the past, whether by policy or custom. Sales will be high in areas with high concentrations of these groups simply because they have been served the longest and are located closest to the outlets. Conversely, some high potential groups may not be present in the trade areas or may be unfavorably located. Sales will be lower to these groups due to their location, not their potential.

Locational problems are particularly important to consider for analysis based on lifestyle clusters. Since any area can be classified in one and only one cluster, it is virtually certain that many clusters, even the majority, will not be represented in a set of trade areas. In this situation it will be impossible to assess the relative favorability of the omitted clusters. Expanding the analysis to include customers who live outside the trade areas only makes things worse. It includes populations from more clusters, but those populations, being outside the trade areas, most likely suffer little "risk" of purchasing the product. The low potential of these clusters will not be one of lifestyles but one of location.

Problems of omission are compounded when zip codes are used for analysis. Since zip codes are larger than census tracts, fewer will be included, seriously limiting the number of clusters under study. Even worse, those that are included may be coded to the wrong cluster. Because of their large size, zip codes often include several neighborhoods belonging to different clusters. Your sales may go to one cluster while the zip code is classified to another. Finally, even though you may have significant sales in a zip code, the majority of that zip code's population may actually live outside of your trade areas. While the potential may be superb, the penetration rate and favorability index will appear moderate or poor. The moral is to use census tracts if possible. They are smaller, and by design more likely to consist of

a single demographic group. The cost of geocoding to the census tract level is modest and should be more than recouped in increased sales.

Locational problems in general demand a more sophisticated multivariate approach to demographic analysis. Distance measures such as travel time (often available from regional planning agencies) or mileage can be used in conjunction with techniques such as multiple regression to control locational factors. Likewise length of time in the market, competitive environment, and store characteristics can be included to control for sales differences resulting from these factors. On the demographic side, using individual demographic variables in conjunction with factor analysis and multiple regression will allow results to be more easily generalized to new markets. If your stores now serve only the "Pools and Station Wagons" cluster, there is no quantitative way through lifestyle analysis to assess the favorability of the "Movers and Shakers" group. On the other hand, high scores on median years of education and median income and low scores on median age may be common to both groups whether or not both are currently within your market areas. The analysis of demographic dimensions in the served area would lead to new targets in areas as yet unserved. The bias of cluster analysis is to target for future service only groups that have been served in the past; the multivariate analysis will be more likely to isolate dimensions characterizing current markets which are also present in potential markets.

Ecological Correlation

In demography, ecological correlation refers to making inferences about individuals based on properties of the social environment in which they live. The concept dates back to studies done in the 1920's which concluded that Jews were more likely than others to commit crimes because neighborhoods with large

Jewish populations had high crime rates. When reanalyzed, the data showed that most of those Jewish neighborhoods had very high levels of poverty. When poverty rates were statistically controlled, Jews were found to have lower crime rates than their neighbors. Poverty not ethnicity led to crime. While similar problems arise in the analysis of survey data, they are particularly severe when using data on the general population. An extensive literature has developed on precisely this point.

Ecological correlation is almost a mainstay of market analysis. This does not invalidate the analysis, but does raise caution flags. A prime example is lifestyle clustering itself, a technique which tends to raise ecological correlation to the level of a social theory. To quote from the blurb on the cover of a recent book on the subject, "your neighborhood speaks volumes about what you eat, drink, drive - even think."[1] That is, you are where you live - the classic definition of the ecological correlation.

Clustering, like all demographic analysis, is a form of data reduction. Hundreds of census variables are reduced to eight or ten dimensions which are then used to identify 40 or more clusters. Some of those clusters may show clear statistical relations with their defining dimensions and offer true insight into the populations which they are meant to describe. Others may be statistical artifacts or even products of the subconscious biases of the statistician who devised them - that is, no more than ecological correlations. In assessing the results of any cluster analysis, first look at the clusters themselves and their relationship to the demographics that define them. Are the demographic index values high? Do they make rational sense? Do they

---

1 Weiss, Michael J. "The Clustering of America", Cambridge: Harper and Row, 1988.

reflect the vendor's characterization of them? Are there alternative explanations of the clusters? For example, do they identify regions of the country like the deep south or the rural Midwest? This can be a sign that the clusters are residual products of the analysis rather than representative of actual demographic groups. Do the clusters identify product usage differences that are not more easily explained by other factors like age or income? Are those differences big enough to be both substantively and statistically significant given the data they are based on? In general, if you have reason to believe that use of your product is related to an underlying demographic dimension like age, income, social status, home ownership, or family structure it is better to use that dimension directly in assessing potential markets or segmenting current customers, if only in the framework of defining the population at risk. The categorization scheme you will derive will be more closely related to the forces that drive your market (and could differ in significant ways from a canned lifestyle analysis). On the other hand, if your product is driven by complex market forces, if its use is highly dependent on image, if the clusters to which it is related are well defined and make sense, you will be better off to use the clustering system and its built in relationships to other products and media habits.

Even more serious than ecological correlation sometimes is "reverse" ecological correlation. This is making an improper inference about the population from what is known about the individual. It is usually the product of inadequately considering all the possible demographic determinants of individual behavior. It can, in some cases, be more harmful than using no analysis at all. Suppose, for example, a company sells health coverage to manufacturing workers. If it were to target zip codes with high percentages in manufacturing it would find many manufacturing workers but they would tend

to be poor, liable to bouts of unemployment, young, and unlikely to live in family households (at least in major urban areas). In general these people cannot or will not buy health coverage. By not considering other dimensions like income, the company would target a very low potential market. In the actual case on which the example is based, percent in manufacturing turned out to have no predictive use even in the complex demographic model that was eventually developed. What may seem simple when looking at individual customers does not necessarily transfer directly to demographic analysis.

The problem of "reverse" ecological correlation can arise when attempting to apply survey results to the population. For the sake of simplicity, assume you have a survey which tabulates sales by age and education. Suppose sales turns out to be determined by the two variables. If the effects are independent, you can use the survey to estimate actual sales very closely if you have a tabulation of sales by education and one of sales by age. You do not need to have the crosstabulation of sales by age and education. On the other hand, suppose age and education are not independent. For example, those aged 65 and above with college educations are twice as likely to buy the product than would be predicted by age and education alone. If you can crosstabulate sales by age and education you can get an almost perfect estimate of actual sales. Sales tabulated by age alone and by sex alone will help, but, depending on the nature of the interaction, may not do a very good job of estimation.

Demographics in most cases consist only of marginals. You have the population by age and the population by education, but not by both variables. Now there are two possibilities for interaction. Not only may age and education not have independent effects on sales, but they will most certainly not be independent in the population. If your sampling scheme was complicated

or did not draw from a universe representative of the population as a whole, then your sample will not show the same relationship between age and education as exists in the data underlying your demographics. If the interactions are serious, a weighting scheme based on a simple use of survey marginals could be highly misleading. There are four solutions:

- Obtain a crosstabulation of the demographic data and apply specific rates estimated from the survey. In our example, obtain the crosstab of the population by age and education and apply age and education specific penetration rates estimated from the survey to each age and education group in the population. Most demographic systems have data crosstabulated by age and sex, age and income, and often age and race which can be used in this way. In particular, age and sex tabulations are extremely useful in any health care application. Use of age and sex specific rates in these areas will almost always outperform other techniques short of having actual population information available.
- Perform more detailed analysis on the survey data itself. Techniques such as loglinear modeling will identify both independent effects and interactions and assess their relative strengths. They can then be used to estimate marginal effects adjusted for the interaction terms. If the interaction terms are not overwhelming, these estimates can be used to construct adjusted rates which can be applied to the demographic marginals. To do this, you must have some confidence that the sampling technique is adequate to reflect the population demographics.
- Geocode the survey responses, append population demographics from each respondent's census tract, and create a model relating the demographics to the survey responses. Since the model is based on

population demographics, it can be applied to population data directly. The technique is relatively expensive and not always possible. But it will produce superior results in any situation where the determinants of buying behavior are complex and actual sales are scattered sparsely over a large geographic area. If proper statistical techniques are used, the method will produce very powerful segmentation models and direct estimates of penetration rates.

- Use the survey only to refine the population at risk. Then use actual sales information in conjunction with lifestyle clusters or demographics. This will work well in situations where you are well established in a market or for analysis of direct mail campaigns where actual response rates are available. In particular, demographic modeling of direct mail response rates will almost always produce significant benefits through target mailing.

Statistical Specification

Statistical specification of demographic models is complex. Major areas you need to consider include:

- Specification of the model. Demographic penetration models are more often than not nonlinear. Penetration can only vary between 0 and 100 and is usually more resistant to change at very high and very low levels. The effect of the independent variables may be reduced as a multiple of distance from a retail site. Logistic transformations of penetration rates work and usually justify their added complexity in the results they get.
- Correlation. All demographic variables are highly correlated both between variables and between locations. Education and income are

always highly related, for example, and neighboring census tracts will be more similar to each other than to those at a distance. These correlations are much higher than those you will find in survey data and will make it extremely difficult to develop a useful regression model. As a first step, then, any demographic modeling effort needs to start with a factor or principal components analysis. The purpose is to reduce the number of independent variables to those representing significant sources of variation in the population statistics. Because of the locational correlation, this analysis will be sensitive to the definition of the market area and may have to be performed each time that definition changes.

- Weighting. Demographic variables are heteroscedastic. That is to say their accuracy varies with the size of the population on which they are based. In a census tract with five people the difference between 20% and 40% is one person. In a tract with 2,500 people, that difference would be based on 500 people. Certainly the data from the second tract will be more stable than those from the first. In particular, models which relate population demographics to individual responses must use adequate weights to account for these differences.

A Final Word

Demographic data available now, no matter how it has been updated, estimated, or projected is ultimately based on the 1980 census. Despite its age, the data provides powerful marketing information when applied correctly. With the coming of the 1990 census this power should take a quantum leap. The technology is in place to make use of the data while it is current; technology that was just being developed when the 1980 census was released. This

technology has created a cheap and abundant source of marketing intelligence. With planning and familiarity with the techniques of analysis, this information can become even more central to the market research process in the next ten years than it has become in the last ten.