Sample Matching Technique: An Approach To Select Better Sample Data

Jigna Mehta

Abstract: Sample survey is very common these days. But projecting the correct population inference using these sample survey information is challenging. Sample weights obtained by weighting procedure does helps in minimizing the sampling bias to an extent but it does not fully eliminates the sampling error. The objectives of this study were how to further refine and use the sample information for population projections prior to putting under weighting techniques.

Key words: Sample, Population, Survey, Weighting, Raking, Benchmarks, Market Research

1 INTRODUCTION

Market research is for discovering what people want, need, believe or act. This type of research helps the business to understand the end consumer behavior. Sample Survey is a mechanism of market research in which questions are asked to a large group of people. These questions aim to measure opinions, attitudes and behavior of those people. One of the survey claims that beer sale increases if they are placed near baby diapers rack. This comes from the understanding that when a father goes for baby products shopping, they tend to buy beer to further reduce their frustration of not hitting the pub. Another mall survey confirms that when bread, butter, egg and other milk products are bundled together, the result in sale is relatively higher. Sample Surveys are not about describing the behavior of respondents in the sample, but about the behavior of the extended group as a whole which is called population. The behavior of the population is summarized in statistics like weighted totals, weighted means and weighted percentages of sample statisitcs. There are two major types of surveys in market research through which data is collected

1. Primary Research – Collected First Hand
2. Secondary Research – Buy the data from vendors

Both of these data collection methods serve the same purpose. It gives us sample data which itself claims to be a good representative of the population of our interest. Once the sample is selected, it has to be treated well so as to make unbiased projections of the population. The most common method is to weight the sample to look like population and then decision numbers are considered. A sample will cover segments of the target population in proportions which may not match the proportions of those segments in the population. One can often improve the relation between the sample and the population by weighting the sample and producing the sample weights so that the marginal totals of the adjusted weights on demographics agree with the corresponding totals for the population. Raking may perhaps reduce nonresponse and noncoverage biases, as well as sampling variability. The sample is usually comparable to the population only on demographic indicators of which detailed population data is available such as age, gender, ethnicity, marital status and income. The weights from the raking process are used in estimation and analysis for population. Weighting produces sample weights. These sample or survey weights are

- A value assigned to each respondent in the sample.
- Normally used to make statistics computed from the sample data more representative of the population.
- The value indicates how much each respondent will count / represent in a statistical analysis such as mean i.e. A weight of 10 means that the respondent represents 10 identical respondents in the population.
- Weights are often fractions, but are always positive and non-zero.

In the computation of means, totals and percentages, not just the values of the variables are used, but weighted values are used. This weighted figure gives an estimate of the population for any attitudinal and behavioral characteristics.

2. POINT AND PURPOSE:

Routinely sample survey followed by weighting is the procedure but this involves time, budgets, extreme weights, instabilities and the resulting weights are not precise due to inconsistent sample and vendor recruitment differences. Hence weighting the sample does not fully eliminates the possible bias which otherwise could have been minimized.
Possible problem of these sample weights could be

1. Weights primarily adjust means and proportions. Reasonable use for descriptive data but may adversely affect inferential data and standard errors.
2. Weights almost always increase the standard errors of the estimates.
3. Extreme weights introduce instabilities.
4. Not allowing the extreme weights further result in reducing the representativeness of the weighted data

Therefore the main objectives of this research were to determine the intermediate solution which could further aid in minimizing the sample data bias prior to producing sample weights. The aim here is to proactively define the well being of the selected sample in terms of its demographics distribution. The idea which was to be tested here was if the sample is good from demographic perspective, then the corresponding behavioral distribution in that of population should not be significantly different.

2. MATERIALS AND METHODS:

2.1 Business Overview

The leading global information and Measurement Research Companies enables other companies to understand consumer's behavior. Such research companies measures and monitors the following

1. What consumers eat (burger, pizzas, pastries),

2. What consumers watch (TV Soaps, Movies, Cartoons, advertising)
3. What consumers buy (brands, products, items)
4. What consumers prefer in technology (internet, text, 3g) etc.. on a global and local basis

2.2 Business Problem

Such companies may have their own data warehouse or they buy data from different vendors, to analyze its objective and then publish the estimated population figures based on the sample figures. In case of owning a data warehouse, there is a constant improvement process and checks in place to incorporate the market changes. But when one has to purchase data from other sources, it's extremely important to check the reliability of the data. Every vendor argues that their data is the best representative of the population. The objective of this paper focuses on which vendor sample data is best. Also, weighting such inaccurate data incurs a high cost and is time consuming too. So the best deal is to have a good, reliable and cost effective sample which further gets refined by weighting to serve the purpose in the best possible way.

2.3 Business Approach

Say a Buyer is using sample data from vendor1 since long time because of its consistency in being representative to the population. So this base data continues to be good in demographics and behavioral characteristics. Now the challenge is to check if other vendor data are as efficient as this base data or not before its gets weighted further. Hence the main attention points are

1. How to check the sample data for its right population representation
2. Reduce the refinement dependency of weighting

3 Methodology and Results

1. Consider the data from another source, say vendor2. Here all the records have equal probability of getting selected
2. Pick up a record randomly from vendor2 data. Match this record's data to the base data only on demographics
3. If a match is found in the base data (vendor1) with respect to demographics, then tag it
4. Every time a record in vendor2 data is selected, its probability of getting selected again is reduced so that another record gets chance for getting selected and compared. Same logic applies to the record in base data. Match new records data to base data on demographics.
5. Repeat the same process until all the records in vendor2 data are selected and compared to the base data
6. Now we have data from vendor2 which is exactly same to the base data on demographics
7. Perform z test statistical testing on behavioral variables between vendor2 and base data and see how much they differ. If there is a significant difference between these behavioral variables, then the vendor2 data is not as efficient as the base data with respect to the behavior.
8. Example:
   a. Base data – Males (50%) and Females (50%)
      Vendor2 data – Males (50%) and Females (50%)
   b. Base data – Use mobile internet (70%), satisfaction (50%)
      Vendor2 data – Use mobile internet (45%), satisfaction (35%)

In the above example, although the demographics are similar in both the data, consumer's behavior towards mobile internet usage and satisfaction differs significantly.

9. Repeat the matching and testing procedure on at least 3 to 6 months of available data so as to further get confirmed on the accuracy of the methodology and vendor2 data
10. Replicate the above process for other different sources of data for different time periods

4 Results and Discussion

An online survey asks respondents for their gender, age and other survey related behavioral questions. Since the distribution of sex and age in the population is known, it is possible to compare the sampling distribution with the population distribution.

	Male	Female	Young	Middle	Old
Population	50%	50%	30%	40%	30%
Sample	51%	49%	60%	30%	10%

The sample consists of 51% males and 49% females. These percentages are close to the corresponding population percentages. So we can say the sample data is a fair representation of population with respect to gender. The sample contains too many young persons and too few elderly. Therefore, the sample data is not a good representative of population with respect to age. We should match the sample data again with other available vendor data. This refined sample should further undergo weighting methodology. Once weights are produced, we perform further statistical testing on how the population might act on the behavioral characteristics.

5 Conclusions

I have tried to give some background on what types of surveys are available and how are they treated further by raking to produce the sample weights. I have also tried to give some background on what are the challenges involved post raking. Hence the idea of an intermediate sample data match helps in minimizing the error prior to raking and further refining the sample before undergoing raking. I have also listed the steps involved in this sample matching technique idea and illustrated with examples on what types of problems can be addressed by this technique

Acknowledgement

I have taken a lot of efforts in testing this idea and putting across the same in paper, however, it would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them. I would also like to express my gratitude towards my family for their

kind co-operation and encouragement which help me in completion of this paper.

REFERENCES

[1] Introduction to Survey Weights by David R. Johnson Department of Sociology, Population Research Institute, The Pennsylvania State University, November 2008

[2] Applied Survey Methods, A Statistical Perspective

[3] Tips and Tricks for raking survey data, by Michael P. Battaglia, David Izrael, David C. Hoaglin, and Martin R. Frankel

[4] Baby Diaper and Beer Survey: http://web.onetel.net.uk/~hibou/Beer%20and%20Nappies.html