Defining the Sampling Frame for the Convenience Store Industry – A Case Study

Eric Falk and Glenn D. White, Jr., Ernst & Young LLP
1225 Connecticut Ave, NW Washington DC 20036

ABSTRACT

Defining the sampling frame and producing estimates for the Convenience Store News Industry Report is a complex problem. With convenience stores opening, closing, merging, and expanding the universe is always changing. Convenience stores are continuously looking for ways to increase revenue and the Industry Report provides competitive intelligence, sales, and marketing trends. This paper will discuss the challenges related to defining the sampling population and for making convenience store industry estimates. In addition to the complicated sample, the survey and publication of the results are done within tight budget and publication deadlines.

Keywords: Sampling Frame, Business Survey

1. INTRODUCTION

Convenience Store News has been conducting a Convenience Story Industry Survey for nearly 25 years. This industry survey focuses on marketing, merchandising, and innovation. In 2004, they expanded the number of categories in the report by teaming with various data partners. They also focused on new areas in the report and used aggregated data. Since 2003 the Quantitative Economics and Statistics (QUEST) group within Ernst & Young has been engaged by *Convenience Store News* to assist with their industry survey. The results of the survey are published in the *Convenience Store News* annual Industry Report.

There is a lot of competition among convenience stores and they are continuously looking for ways to increase revenue. The Industry Report published by *Convenience Store News* is a useful source of information for the stores to use in trying to plan strategies to reach their goals.

Defining the sampling frame and producing estimates for the convenience store industry is a complex problem. With stores opening and closing the universe is always changing. Defining the sampling unit is also a difficult task, for example, what ownership structure defines a multiple store versus a single store?

2. BACKGROUND AND OBJECTIVES

One of the goals of this study is to make industry estimates for the convenience store population in the United States. According to the *National Association of Convenience Stores (NACS)*, a convenience store has to have:

1) Size of less than 5,000 sq ft;
2) Off-street parking;
3) Extended hours;
4) Offer at least 500 stock keeping units (SKUs); and
5) A product mix including grocery type items.

QUEST has been involved in the project for two years and is responsible for the survey instrument design, the survey implementation and management, the sampling and estimation. The estimates made for the Industry Report are a combination of data gathered from three sources: 1) respondents from a survey sent to a sample of convenience stores, 2) industry data, and 3) government data

A goal of the Industry Report is to provide information on in-store and out-of-store sales for convenience stores. The in-store sales include: 1) merchandise, such as cigarettes and alcohol and 2) food service, such as deli meats and sandwiches. The out-of-store sales include only motor fuel.

Another goal of the Industry Report is to show whether there is a difference between the single stores and the big chain stores such as 7-Eleven, Wawa, and Sheetz to name a few. The ultimate challenge encountered during this study was getting reliable category data for in-store sales. For example, in-store merchandise sales, attempts to estimate the total sales of cigarettes, beer & malt beverages, wine & liquor and candy & gum. Due to low response rates these percentages were difficult to accurately estimate.

Another example of category type data is the method of payment; for example, cash, credit cards and debit cards. This information is not available through any alternative source and we had to survey the convenience stores to collect this data.

3. CHALLENGES

These were several challenges related to the sampling and estimation portion of the convenience store industry project.

3.1   One of the sampling challenges was defining the sampling population. The universe of convenience stores is constantly changing with new stores opening and old stores closing. In addition, there were some stores that were not considered to be true convenience stores based on the NACS definition of a convenience store. Also, there were some miscellaneous stores that were removed manually when defining the sampling population.

3.2   The second sampling challenge was defining the sampling unit. For purposes of this study, the definition of a single store was a convenience store not associated with any other store. The definition of a multiple store was a convenience store associated with at least one other store. The association of the cluster of stores was made through the owner of the stores or ultimate parent company.

3.3   The third sampling challenge was related to the distribution of the number of stores associated with the ultimate parent companies. The range of stores associated with multiple stores ranged from two to several thousand as in Wawa, Sheetz and 7-Eleven. Thus, the multiple store population was a highly skewed distribution.

3.4   The fourth sampling and estimation challenge were the tight deadlines and a fixed budget to design and select the sample and produce the estimates. The deadline for selecting a sample was less than two weeks from the time that we received the population data. In addition, we had about two weeks to produce fifty tables of estimates for the Industry Report when data collection closed.

3.5   The fifth sampling and estimation challenge was that the estimation was complicated by low response rates for the category data and we needed to incorporate data from other sources to stabilize the estimates.

4. SAMPLING POPULATION

The full population of convenience stores was provided electronically by Trade Dimensions (TDLinx). This data set is considered by *Convenience Stores News* and *NACS* to be the industry standard. The data were provided to us in November 2004 and represented a snapshot of convenience stores in the United States.

Some of the fields included in the dataset were 1) name and street address, 2) ultimate parent company, 3) store size and 4) company size.

The database contained 137,096 convenience stores based on the TDLinx definition of a convenience store. The database is constantly changing from year to year and stores are commonly identified and added to the list by suppliers. During a recent TDLinx audit several years back, 8,000 stores were discovered using the supplier distribution lists.

The database was contained some inconsistencies. For example, there were many instances where the address was not correct or current, the owner information was not correct or current and some of the key information was not correct. In addition, there were stores we needed to remove because they were considered out–of-scope. This included smoke shops (these are stores that primarily sell tobacco products), closed stores, duplicate stores or because of inconsistent definitional issues. After removing the out-of-scope stores the final sampling population was 137,117 convenience stores.

5. SAMPLING POPULATION SUMMARY

Table 1 summarizes the convenience stores in the TDLinx database and the number that were removed to define the final sampling population. For example, to identify the convenience stores that were smoke shops we listed any convenience store with the character string "Smoke", "Tobacco", "Smoke Shops" etc. These were reviewed manually and removed when it was apparent that the store sold mainly tobacco products.

Table 1 - Sampling Population

Description	# of Stores
Full TDLinx Population	**137,906**
- Smoke Shops	25
- Closed Stores	264
- Duplicate Stores	260
Defintions	240
Final Sampling Population	**137,117**

6. DEFINING THE SAMPLING UNIT

One of the objective of our analysis was to provide industry information for single stores, multiple stores, and overall. Single stores were defined as convenience stores that were "mom and pop" type stores, not associated with any other convenience stores. Multiple stores, as mentioned, were defined as convenience stores that were associated with at least one other convenience store. For example, there were over 5,700 convenience stores associated with one chain.

In initial discussions the sampling unit was considered to be the individual convenience store. However, for multiple stores there was a higher likelihood of getting information regarding a convenience store through the owner or ultimate parent company. In fact, the owner of the chain of stores was able to provide information regarding all of their stores that they owned or operated. Thus, the sampling unit became the individual owner. For the single stores, the number of stores associated with the owner was one and for the multiple stores, the number of stores associated with the owner was two or more.

The sampling unit for the multiple stores was done by aggregating the individual stores to an ultimate parent company. The process was difficult mainly because of definitional issues. The definition used by the developers of the sampling frame, TDLinx, changed from last year to this year and led to some unforeseen circumstances. For example, stores in two large chains appeared under the definition as independent single stores, but in fact they should be placed under the chain headquarters. Some of the individual stores needed to be reviewed manually and recorded when this occurred. In addition, there were inconsistent fields within the TDLinx database, such as owner information and company size.

7. SAMPLE DESIGN

After applying the methodology for single and multiple stores there were 81,686 single stores and 55,431 stores associated with 1,877 owner/ultimate parent companies. The number of stores associated with the owner of the multiple stores ranged from two to several thousand.

A consideration for the sample design was that we expected low response rates given our previous experience with the study. Because of the fear of very small strata and the relative homogeneity in store sizes among the single stores and small chains, we designed only two strata for the single stores and two strata for

multiple stores; a certainty stratum and non-certainty stratum.

8. SAMPLE DESIGN – OVERALL

Table 2 shows the overall sample design and sample selected per stratum. Stratification was by single and multiple stores and then within each of these strata there was a certainty and non-certainty stratum.

Single stores with average weekly dollar sales of $90,000 or more were selected in the certainty stratum. The total sample size for the single stores was 400. For the multiple stores, any owner/ultimate parent company that was associated with 70 or more stores was selected in the certainty stratum. The sample for multiple stores was also 400.

Table 2 – Sample Design

Single Stores

Stratum	Definition	# of Stores in Population	# of Stores in Sample
1	Weekly Volume < $90K	81,636	350
2	Weekly Volume >=$90K	50	50
Total		81,686	400

Multiple Stores

Stratum	Definition	# of Ultimate Parents in Population	# of Ultimate Parents in Sample
3	# of Stores Associated < 70	1778	301
4	# of Stores Associated >= 70	99	99
Total		1,877	400

9. NEXT YEAR

In preparation for the study next year we anticipate that the creating of the sampling frame should take less time. This assumes that the definition does not change and that that the input files provided by TDLinx will be consistent.

For next year and years going forward we are considering setting up a panel of 30 to 50 multiple stores that would provide us with the elusive category information and our hope is that this base along with the sample will help strengthen the final estimates. We plan to select these stores as part of the random sample using a permanent random number which would ensure a higher probability of selecting the same stores from year to year.

ACKNOWLEDGEMENTS