

Introduction

The Synar sample is drawn from a sampling list created from the Department of Revenue's Cigarette License File (CLF), which contains the name and address of every outlet that purchased a license to sell cigarettes in Pennsylvania. Since the survey uses a list to draw the sample, Substance Abuse and Mental Health Services Administration (SAMHSA) requires that Pennsylvania test that list. Pennsylvania conducted a coverage survey in 2013 to test the quality of the sampling frame list and fulfill the federal requirements.

Results

In 2013, Pennsylvania carried out a coverage survey that resulted in a weighted coverage rate of 99 percent. Pennsylvania estimates that 97-100 percent of outlets that sell cigarettes in Pennsylvania are accounted for on the sampling frame which, is derived from the Department of Revenue's CLF.

The survey was conducted in the month of June by Department of Drug and Alcohol Program employees. The surveyors entered an outlet and determined if that outlet sold cigarettes. If cigarettes were being sold, the name and address of the establishment was recorded.

The addresses collected during the survey are matched to the current Synar sampling frame. Both electronic and manual matching techniques are used. As shown in Table 1, each address collected from the survey has one of five possibilities:

1. **Exact match:** The address exactly matched an address on the sampling frame.
2. **Close match:** The address closely matched one on the sampling frame but had a small, insignificant difference. For example, 32 4th street is considered a close match to 32 Fourth Street.
3. **Not enough info:** There was not enough information to make a determination.
4. **No match:** The address collected was not an address on the sampling frame.
5. **Non-completion:** The address was not completed by the surveyor or was outside of surveying area.

Table 1. Initial and unweighted analysis of survey data

Coverage Type	Frequency	Percent
Close	80	47.1
Exact	87	51.2
Not enough info	0	0.0
No match	1	0.6
Non-completion	2	1.2
TOTAL	170	

Table 2 displays results for statewide and individual strata. The statewide total includes all strata. The weighted statewide coverage rate was 97-100 percent. The state stratum, defined in the design and methodology section, had a weighted coverage rate of 100 percent. The remaining categories, Allegheny and Philadelphia counties, are displayed in the table, the weighted rates are not displayed because the design of the survey does not allow for county estimates.

Table 2. Results by area

Area	Projected Outlets	Outlets Completed	Outlets Matched	Outlets Unmatched	Wgtd Match Rate (%)	Lower Limit (%)	Upper Limit (%)
Statewide total	170	168	167	1	99	97	100
State stratum	140	138	138	0	100	100	100
Allegheny county	10	10	9	1	n/a	n/a	n/a
Philadelphia county	20	20	20	0	n/a	n/a	n/a

Design and Methodology

A valid survey relies on a valid sampling frame. Assessing the quality of the sampling frame is an important part of the survey procedure. When assessing the quality of a list-based sampling frame, two deficiencies are usually measured, over-coverage and under-coverage. Over-coverage occurs when the frame contains ineligible outlets, such as outlets that are not accessible to youth, duplicate addresses, outlets no longer open for business or outlets that no longer sell cigarettes. The over-coverage rate is calculated with the Synar data and does not require a separate survey. Under-coverage occurs when eligible outlets are not included on the frame, a much more serious deficiency than over-coverage. Under-coverage requires a separate survey. The purpose of this coverage survey is to estimate the under-coverage of the Pennsylvania Synar sampling frame.

The Center for Substance Abuse Prevention (CSAP) oversees each state's implementation of the Synar amendment and they require each state to "develop a sampling frame that includes, at a minimum, 80 percent of the tobacco outlets in the state." The 2013 coverage survey provided the data necessary to fulfill CSAP's requirement to report Pennsylvania coverage rates.

The state is divided into three mutually exclusive geographical strata: Allegheny county, Philadelphia county and the rest of the state (state stratum). The state stratum is divided into mutually exclusive clusters consisting of zip codes or groups of zip codes. These clusters are the primary sampling units (PSU) for the survey. It is estimated that each PSU contains roughly 50 to 200 outlets. Allegheny and Philadelphia are not clustered.

A two-stage sampling method is used in the state stratum. During the first stage, 14 PSUs are randomly selected using probability proportional to size sampling, where the size of the PSU refers to the estimated number of outlets within that PSU.

During stage two, one "starting address" is randomly selected from each selected PSU. The "starting address" is determined by randomly selecting an address from the Census Bureau's Tiger file and then entering that address into Google Maps to find the nearest convenience store to that address within the sampled PSU. If Google Maps does not recognize the address, then Google Maps uses the zip code centroid; the nearest convenience store to the centroid is used as the "starting address."

The surveyor records the location information for each randomly selected starting outlet and then follows a pre-determined route to canvass the area for nine more outlets.

The survey routes are created using the knowledge that almost every zip code has a post office. First, every zip code in the selected PSU or county is randomly sorted. Next, a route from the starting point to the post office of the first randomly ordered zip code in the PSU or county is mapped using an online mapping system (Google Maps was used for directions and the options "Avoid Highways" and "Avoid Tolls" were enabled). Next, a route is created from the post office of the first zip code to the post office of the second zip code, from the second to the third and so on. If a zip code does not have a post office, the zip code centroid is used to create the route. Since the starting point is randomly selected and the post offices are randomly ordered, these are random routes.

The methodology for the Allegheny and Philadelphia strata is slightly different than the state stratum. Allegheny and Philadelphia have one stage of sampling. One starting point is randomly selected in Allegheny and two starting points are randomly selected in Philadelphia using the Tiger file. The same procedure as above is then followed to select the starting convenience stores and the random routes the surveyors will follow to locate the remaining nine outlets for each starting point.

Coverage Survey Weighting

The base (design) weight is the inverse of the probability of selection. The probability of selecting an outlet is calculated from the probability of selecting the nearest address. For the random strata, the probability of selecting any address is the total sample size of the stratum divided by the total population of the stratum.

The probability of selecting an address in the state stratum (clustered) is a little more complex. The state stratum employs a cluster design where the clusters are selected using probability proportionate to size (PPS) sampling. The overall probability of selecting an outlet is the product of the probabilities of each stage of sampling (the probability of selecting a cluster in the first stage, multiplied by the probability of selecting an address in the second stage).

Let,

PROBCL = Probability of selecting a cluster

PROBADD = Probability of selecting an address within the cluster

PROBST = Probability of selecting the address in the stratum

NCLUST = Number of clusters in the stratum

CPS = Cluster population size (cancels out in PPS sampling)

ELIGN = Eligible stratum population size

SAMPSIZE = Sample size of the cluster

$$(i) \quad \text{PROBCL} = (NCLUST) \times \left(\frac{CPS}{ELIGN} \right)$$

$$(ii) \quad \text{PROBADD} = \left(\frac{SAMPSIZE}{CPS} \right)$$

$$(iii) \quad \text{PROBST} = (\text{PROBCL}) \times (\text{PROBADD})$$

$$(iv) \quad \text{PROBST} = (NCLUST) \times \left(\frac{CPS}{ELIGN} \right) \times \left(\frac{SAMPSIZE}{CPS} \right)$$

$$(v) \quad \text{PROBST} = (NCLUST) \times \left(\frac{SAMPSIZE}{ELIGN} \right)$$

$$(vi) \quad \text{Base (design) Weight} = \left(\frac{(ELIGN)}{(NCLUST)(SAMPSIZE)} \right)$$

The final weight adjusts for non-completion. Let, FINALWT = final weight and SAMPOBS = the number of completed and eligible sample per cluster.

$$\text{Final Weight} = \text{Base Weight} \times \left(\frac{SAMPSIZE}{SAMPOBS} \right)$$

The Bureau of Health Statistics and Research welcomes comments and suggestions on the content and format of this report. Staff is available to answer questions regarding the report, including utilization and limitations of the data. Please address all comments and questions to:

Pennsylvania Department of Health
Bureau of Health Statistics and Research
555 Walnut Street, 6th floor
Harrisburg, PA 17101
Or call: 717-783-2548