PA Synar Survey Design

The purpose of most surveys is to make an inference about an entire population by only examining a small sample of the population. The Synar survey estimates the rate that outlets sell cigarettes to minors and uses a confidence interval to make an inference about the population of cigarette selling outlets. The sample used for this inference is a probability sample where every outlet in the population has a non-zero probability of being selected into the sample. By adhering to this condition, statistical inference techniques can be used to make inferences about the population.

The survey was developed using a stratified and clustered design with the primary purpose to limit the amount of error at the state level. Other levels of estimates can be obtained but they will have much larger amounts of error. For example, valid statistics can be obtained for the Health Districts, Allegheny County, Delaware County, Erie County and Philadelphia County but there will be a large amount of error. Estimates are not valid at the county level (other than the four counties mentioned above). There is a common misconception that the data collected for a particular county (other than Allegheny, Delaware, Erie or Philadelphia) can be used to obtain a "rough" estimate of the county's violation rate. This is not true unless the entire county is a cluster. In fact, county level statistics are more misleading than making up numbers based on an educated guess. Please do not use the data in this way.

> ⚠ *Important: Due to survey design, results CANNOT be calculated for individual counties other than Allegheny, Delaware, Erie and Philadelphia.*

> ⚠ *Enforcement activities are not combined with the survey. The survey is used for measurement purposes only.*

The SAMHSA Synar guidelines allow vending machines to be eliminated from the survey if they are not located in places accessible to youth. Act 112 validates the removal of vending machines from the survey but if a sampled outlet is selling only through a vending machine, it must be attempted.

Sampling Frame

A sampling frame is used to identify every element in a population. Two popular types of frames are the area frame and the list frame. An area frame contains information for geographical areas while the list frame contains identification information for the actual population elements. The Synar survey uses a list frame.

The Synar population is defined as "every outlet in Pennsylvania that sells cigarettes and is accessible to minors", therefore the sampling frame must contain the name and address of every outlet that sells cigarettes and is accessible to minors. Most of this information can be obtained 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 PA. The sampling frame is created from the CLF. According to Pennsylvania law, each retailer must purchase a license and provide the name and address of the location before they can sell cigarettes from that location. The only deficiency of using the CLF is that it won't capture an outlet that never bought a license. If the outlet

bought at least one license, it will be in the CLF database. Outlets selling cigarettes without a license are part of the Synar target population. Outlets could sell cigarettes without a license or they could purchase a license after the frame is created. There are two methods used to counteract these problems. First, the Department of Revenue actively identifies and prosecutes outlets that sell cigarettes without a license and therefore limits the number of outlets willing to sell illegally. Secondly, viable locations are left on the file even after they claim to be out-of-business or no longer selling in case they decide to start selling illegally or after the frame is created.

> ⚠ *The Synar sampling frame is used to identify outlets that sell cigarettes, NOT outlets that purchased a license.*

The quality of the sampling frame is important to the survey processes. There are two main types of deficiencies to measure when assessing the quality of a list based sampling frame like the Synar sampling frame; 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, private clubs or private residences. Over-coverage is not a serious threat to the validity of the survey and there are ways to prevent it before the frame is created. For example, a series of computer programs are used to identify and eliminate duplicates, unusable records and ineligible outlets (e.g., VFW's, Elk clubs, etc.) before the frame is created. The sample frame review procedure (discussed later) is another way that identifies ineligible outlets on the Synar frame.

Although many steps are taken to prevent over-coverage, there will always be some instances of over-coverage to be dealt with during the field work. Field representatives should expect over-coverage. Some outlets are kept on the list even though they seem to be ineligible. It is better to keep an ineligible outlet on the list than to remove an eligible one by accident. To protect the validity of the survey, outlets that haven't sold cigarettes or have been out-of-business cannot be automatically taken off the current list unless they have been visited in the last 12 months. It may seem like an inconvenience but we are trying to produce the best results possible. With few exceptions, if the dwelling of the outlet remains in workable condition then it is kept on the list to prevent against the purchase of a cigarette license after the frame is created or selling without a license. To maintain survey integrity, these outlets cannot be taken off the list because our population is not defined as those outlets that buy a license but rather those outlets that sell cigarettes, either legally or illegally. It is much more harmful to the survey to eliminate outlets that should be on the list than to have outlets on the list that no longer sell cigarettes or out-of-business.

Under-coverage, on the other hand, is a serious quality problem and can be difficult to handle. Under-coverage is a term used to describe a frame that is missing eligible outlets. The Department of Revenue minimizes under-coverage by enforcing the law that every outlet selling cigarettes must purchase a license. By actively enforcing the license law, the number of outlets selling cigarettes without a license and not on the frame will be at a minimum, but there will always be those that choose to operate outside of the law. There are many reasons why an eligible outlet may not appear on the list. Reasons for not being on the list can range from the outlet is selling cigarettes without a license to the address was incorrectly entered into the database. All of these are under-coverage issues and can bias the survey results. If an outlet is eligible and not on the list, it doesn't have a chance to be selected into the survey and it has a zero probability of selection. A probability survey, like the Synar survey, is based on the assumption that every element of the population has a non-

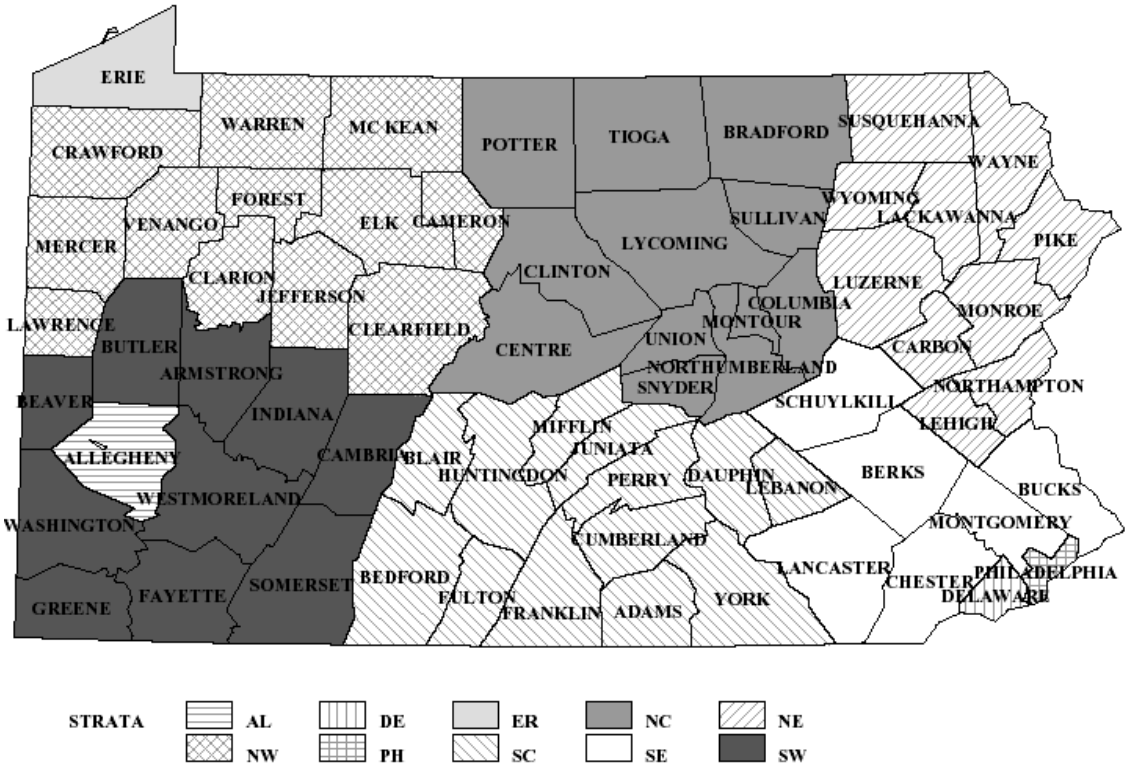
zero probability of selection and the zero probability caused by under-coverage will jeopardize the survey results. If the missing outlets all had a common trait then an entire section of the population that doesn't have a chance to be selected into the survey because they are not on the list. Imagine further if this group had the same preference on selling cigarettes to minors.

Under-coverage is such a serious threat to survey validity that CSAP requires the testing of the sampling frame every few years. The list is tested using a coverage survey designed to check for eligible outlets not on the list. If the coverage rate (rate of eligible outlets on the list) is too low, another way to create the sampling frame will have to be devised.

Stratification

The entire population is placed into separate and distinct subpopulations or strata (*see* **Figure 13**). Every eligible outlet location on the sampling frame is grouped into 10 mutually exclusive and exhaustive geographical strata consisting of the Northcentral Health District (NC), Northeast Health District (NE), Northwest Health District (NW), Southcentral Health District (SC), Southeast Health District (SE), Southwest Health District (SW), Allegheny (AL), Delaware (DE), Erie (ER) and Philadelphia (PH). If the survey error is not too high, data collected for these strata can be used to develop separate within-stratum estimates and statewide estimates after weighting.

Figure 13. 2015 Synar Strata



Clustering

The outlets within the six "District" strata (NC, NE, NW, SC, SE and SW) are grouped into

geographic clusters of adjacent zip codes. A predetermined number of these clusters are sampled and a predetermined number of outlets are selected within the cluster. Since the outlets in each cluster are designed to be geographically close, travel costs are reduced. The size (number of outlets) of the cluster can vary between 40 and 1 less than the sampling interval of the PSU. Every year new outlets are added, old outlets are removed and clusters are reviewed and adjusted so they comply with the size criteria. The outlets in Allegheny, Delaware, Erie and Philadelphia counties are not part of the six District strata, they are referred to as the random areas or strata.

Sampling

Clustered Areas

The clustered areas use a two stage sampling design. The first stage consists of selecting a pre-determined number of PSUs (clusters) from within each stratum using the probability proportionate to the size (PPS) sampling technique. Size of the cluster refers to the number of outlets within the cluster. PPS sampling assures that the larger clusters (more outlets) have a greater chance to be selected but the probability of selection for each individual outlet remains equal. Figure 14 illustrates the first stage sampling design.

Stage two involves randomly selecting a pre-determined number of outlets from each of the sampled clusters. PPS sampling demands that the same number of outlets are selected from each cluster. The current sampling plan requires 17 outlets to be selected from each cluster that is sampled.

Random Areas

The outlets within the single county strata of Allegheny, Delaware, Erie and Philadelphia are selected using a simple random selection process. For each county stratum, every outlet is given a random number, sorted by that random number and a pre-determined number of outlets are selected from within each of the counties.

Figure 14. First Stage Sample Design