Creating a Flexible and Scalable PSU Sample for NHTSA's Redesign of the National Automotive Sampling System

Martha Rozsi[1], William Cecere[1], Sharon Lohr[1], James Green[1]
[1]1600 Research Boulevard, Rockville, Maryland 20850

Abstract

The redesign of the National Automotive Sampling System (NASS) required a flexible and scalable PSU sample to be able to respond to future and changing budget levels and precision needs. It was assumed that the future sample for NASS could have between 16 primary sampling units (PSUs) and 101 PSUs for the Crash Report Sampling System module, and between 16 PSUs and 96 PSUs for the Crash Investigation Sampling System module. This paper describes an approach that allows the number of PSU strata, and thus depth of stratification, to change in response to changes in budget and thus total PSU sample sizes. Conditional probabilities are calculated that allow the PSUs in the largest PSU sample to be subsampled as needed to meet future budgetary levels.

Key Words: Primary sampling units, sample design, stratification, scalable sample size

1. Background

The National Highway and Traffic Safety Administration (NHTSA) currently administers the National Automotive Sampling System to sample motor vehicle crashes. There are currently two parts to the NASS: the General Estimates System (GES) and the Crashworthiness Data System (CDS). The GES is a national representative sample consisting of 60 PSUs that provides national estimates of passenger vehicle crashes by sampling police accident reports (PARs). The CDS consists of 24 PSUs sampled from the GES's 60 PSUs. In addition to collecting information from PARs, the CDS also conducts interviews with those involved in the crashes, as well as takes extensive measurements of the vehicles and the sites at which the crashes occurred.

The original design of the NASS was conducted in the late 1970's, with a redesign done in the late 1980's. The NASS needed to be updated to reflect changes since the last design which include:

- Changes in the scope of the project, such as adding additional modules besides the GES and the CDS
- Changing the target populations and sample allocations to reflect oversampling the target populations
- Making the GES and the CDS independent of one another (an option especially of interest as electronic PARs (or ePARs) become increasingly available)
- Changes in the population of the United States
- Changes in the budget (which requires a flexible and scalable PSU sample)

In the course of the redesign, the GES and the CDS were subsequently renamed to the Crash Report Sampling System (CRSS) and the Crash Investigation Sampling System (CISS), respectively.

The future budget is currently unknown, and even if it was certain at this point in time, could possibly change overtime. Therefore, a scalability plan that would allow the number of PSUs in the sample to change in response to resources was derived. This scalable plan would need to subsample PSUs if the budget contracts, but also allow to add PSUs if the budget expands. This scalability plan also requires that the correct probabilities of selection are maintained as these are needed for unbiased estimation.

2. Initial PSU Sample Designs

In order to proceed, both a minimum and maximum PSU sample size were assumed for each of the CRSS and CISS. It was also assumed that, generally, 2-PSU-per-stratum designs were preferable for both the CRSS and CISS.

2.1 Initial Maximum and Minimum PSU Sample Sizes

The first step in developing a scalable design was to determine the maximum number of PSUs that could possibly be implemented. These maximum PSU sample sizes were determined by doubling the current budget, and therefore doubling the current PSU sample sizes of 60 and 24. The CRSS was determined to only have a maximum sample size of approximately 100 PSUs instead of 120 PSUs as this study is to produce national estimates using only information gathered from PARs, and therefore would likely not need more than 100 PSUs to accomplish most reasonable precision objectives. Due to certainty PSUs when sampling with a probability proportional to size (PPS), the maximum number of PSUs for CRSS was set to 101 PSUs (1 certainty PSU), and the maximum number of PSUs for CISS was set to 49 PSUs (1 certainty PSU). The minimum number of PSUs for both CRSS and CISS were determined to be 16 each, based on a shrinking budget but also requiring a minimum number of cases to provide estimates.

2.2 PSU Stratification

The initial stratification started with 8 major strata determined by combining the 4 US Census regions and 2 urban-rural classifications for the PSUs. These 8 major strata are considered the basis of the stratification for the scalability, and are used for the minimum sample size of the 16 PSUs (sampling 2 PSUs per stratum).

Within these 8 major strata, additional PSU strata were created to reach the 50 noncertainty strata for CRSS and the 24 noncertainty strata for CISS with a view to minimizing variance for specified variables of interest. Variables related to vehicle crash types of interest were used to form the strata, including vehicle miles traveled (VMT), fatal crash counts, total crashes, large truck mileage (for CRSS only), and miles of road by road type. A hierarchical process was used to form the substrata based on these variables. The process began by splitting each major stratum into two substrata using one of the stratification variables. Then, each of the two substrata was further divided using additional stratification variables. The end result was a tree describing the nested substrata (Krenzke and Haung, 2009).

2.3 Creating Scenarios Based on PSU Sample Sizes

The initial strata (50 strata for CRSS and 24 strata for CISS) and maximum number of PSUs are considered to be Scenario 1. To then allow for a flexible design that could go from 101 PSUs to 16 PSUs for CRSS, various intermediate points were chosen to define other specific scenarios. The plan was to have 5 scenarios for both CRSS and CISS, where the scenario determines the PSU sample size and PSU stratification. Ultimately, there were 5 scenarios for CRSS, but additional scenarios were added to CISS. The additional scenarios for CISS were added with the understanding the CISS itself needs to be flexible, as it could be used for additional studies in the future, and therefore could require additional PSUs, i.e., beyond the maximum number previously agreed upon. Therefore, the number of scenarios for CISS increased from 5 to 7.

For CRSS, there are 5 PSU sample size scenarios, ranging from the 101 PSU sample for Scenario 1 to the 16 PSU sample for Scenario 5, with evenly spaced intervals in-between for Scenarios 4, 3, and 2. Scenarios 1, 2, and 3 for CRSS all have a certainty PSU, and therefore the number of CRSS PSUs to sample is odd as the certainty PSU is in its own stratum.

Similarly, for CISS, there were originally 5 PSU sample size scenarios, ranging from the 49 PSU sample for Scenario 1 to the 16 PSU sample for Scenario 5, with evenly spaced intervals in-between for Scenarios 4, 3, and 2. For CISS, only Scenario 1 contains a certainty PSU. After these 5 initial scenarios were created, two additional scenarios were added to possibly double the sample size of the 48 noncertainty PSUs. Due to additional certainties in the added scenarios, the new maximum PSU sample size for CISS is 96 PSUs. Because the increase in CISS sample size occurred after stratification and sampling of the initial 49 PSUs, the 2 added scenarios are considered a 3 PSU per stratum design and a 4 PSU per stratum design using the original 24 strata discussed previously.

Table 1 shows the PSU sample sizes and number of noncertainty strata for each scenario for both the CRSS and the CISS.

Table 1: PSU Sample Sizes by Scenario for CRSS and CISS

	CRSS		*CISS*	
Scenario	*Number of Noncertainty Strata*	*Number of PSUs*	*Number of Noncertainty Strata*	*Number of PSUs*
0	N/A	N/A	24	96
0.5	N/A	N/A	24	73
1*	50*	101*	24*	49*
2	37	75	20	40
3	25	51	16	32
4	12	24	12	24
5	8	16	8	16

*Indicates initial designs

3. Scaling from Scenario to Scenario

3.1 Determining the Number of PSU Strata for Scenarios 2 through 5

The goal was to create a flexible PSU design while maintaining a high degree of stratification for each scenario. The smallest design, Scenario 5, will use the 8 major strata based on Census region and urbanicity. The strata for each scenario were allocated based on the total PSU measure of size (MOS) for each of the 8 major strata, while also forcing a minimum of 2 strata within each of the 8 major strata for Scenarios 1 through 3, and 1 stratum per major stratum for Scenarios 4 and 5. This was done to ensure all PSUs have a chance of selection while still trying to balance the population proportionally. Table 2 shows the allocation of PSUs per major stratum for Scenarios 1 through 5 for CRSS and CISS.

Table 2: Allocation of Strata by Scenario for CRSS and CISS

	Major Stratum			*Scenario*		
		1	2	3	4	5
CRSS	1	8	6	4	2	1
	2	2	2	2	1	1
	3	7	6	3	1	1
	4	4	3	2	1	1
	5	14	10	6	3	1
	6	6	4	3	1	1
	7	7	5	3	2	1
	8	2	2	2	1	1
CISS	1	3	3	2	2	1
	2	2	2	2	1	1
	3	4	2	2	2	1
	4	2	2	2	1	1
	5	6	5	2	2	1
	6	2	2	2	1	1
	7	4	2	2	2	1
	8	2	2	2	1	1

3.2 Determining The PSU Stratification for Scenarios

Based on the allocation of strata for each scenario, the next step in getting the flexible design was to collapse the strata in Scenario 1 for both CRSS and CISS to the number of strata allocated in Scenario 2. This ensured a properly nested design while respecting PSU allocation based on the population proportions. Once the Scenario 2 strata were determined, then the Scenario 2 strata were collapsed to get the Scenario 3 strata using the same approach. Likewise, the process was repeated to get to the Scenario 4 stratification. Scenario 5 stratification did not need to follow this process as it was already determined that Scenario 5's stratification would use the 8 major strata to get the minimum 16 PSU designs. To determine the stratification for each scenario, a series of steps were completed to determine the optimal stratum collapsing.

3.2.1 Step 1 – Review Certainties

As stated in Section 2, a certainty PSU existed in Scenarios 1 through 3 for CRSS as well as Scenario 1 for CISS. The first step was to determine if the PSU was still a certainty in the next scenario based on the number of PSUs to be sampled for that scenario and its PSU stratification. If the PSU was not a certainty, such as in Scenario 2 for CISS, then the PSU was placed in the stratum in which it would belong in based on its VMT, fatal crash counts, total crashes, large truck mileage (for CRSS only), and miles of road by road type, which were used to determine the stratification in the previous scenario. If the PSU was still a certainty, as in Scenario 2 for CRSS, then the PSU was in its own stratum and was removed from the next calculations for determining stratum definitions.

3.2.2 Step 2 – Collapse Strata

The next step was to collapse noncertainty strata so that the number of strata per major stratum matches the number of PSU strata determined in Table 2. The hierarchical stratification structure described in Section 2.2 was exploited to collapse strata. For example, major stratum 3 in CISS Scenario 1 had four secondary strata, formed by splitting first on VMT and then splitting the two substrata on total crashes and road type, respectively. These four secondary strata were consolidated into two secondary strata by removing the splits on total crashes and road type, i.e., splitting the major stratum only on VMT. The hierarchical stratification procedure used attempts to keep the total MOS for the substrata approximately equal at each stage in the stratum construction process, so this collapsing method resulted in collapsed strata with approximately equal MOS's (Krenzke and Haung, 2009).

3.2.3 Step 3 – Repeat Steps for Each Scenario

Before sampling of PSUs could be done, the PSU stratification was determined for each scenario for both CRSS and CISS. Therefore, the collapsing of strata was repeated to get the strata definitions for all 5 scenarios for both CRSS and CISS. These collapsed strata were then used to calculate the probabilities of selection for each PSU, and these probabilities were used for sample selection.

4. Calculating PSU Probabilities of Selection

Both the conditional and unconditional probabilities of selection were calculated for each PSU to ensure unbiased estimation.

4.1 Calculating the Conditional PSU Probabilities of Selection

The conditional PSU probabilities of selection were calculated knowing the previous scenario (i.e. Scenario *n*) stratum and the future scenario (i.e. Scenario *n+1*) stratum. These conditional probabilities were used as the scenario-specific PSU selection probabilities.

Let the original probability of selection for a given PSU in Scenario 1, PSU stratum A be determined as follows:

$$p_{1,A} = \frac{n_A MOS_i}{MOS_A} \qquad (1)$$

where n_A is the number of PSUs selected from stratum A and MOS_A is the total MOS for all PSUs in stratum A. Similarly, let the original probabilityof selection for a given PSU in Scenario 1, PSU stratum B be determined as follows:

$$p_{1,B} = \frac{n_B MOS_i}{MOS_B} \qquad (2)$$

where n_B is the number of PSUs selected from stratum B and MOS_B is the total MOS for all PSUs in stratum B.

The conditional probability of selection for a given PSU in Scenario 2, collapsed PSU stratum A and B is:

$$p_{2|A} = \frac{MOS_A}{MOS_A + MOS_B} \times \frac{1}{n_{1,A}} \times n_{2,A\cap B} \qquad (3)$$

where $n_{1,A}$ is the number of PSUs sampled in the previous scenario (in this case Scenario 1) stratum A and $n_{2,A\cap B}$ is the number of PSUs to sample from this stratum in the next scenario (in this case Scenario 2). In this procedure, $n_{1,A}$ will equal 2 unless the stratum that had certainties in the previous scenario no longer has a certainty, in which case $n_{1,A}$ will then equal 3. Also, $n_{2,A\cap B}$ will equal 2 except for in between scenario sizes, such as wanting a sample of 60 PSUs which is between Scenarios 2 and 3, in which case $n_{2,A\cap B}$ could take on values of 2, 3, or 4.

4.2 Calculating the Unconditional Probabilities of Selection for a PSU

The overall unconditional probability of selection for a given PSU in the next scenario, collapsed PSU stratum A and B is therefore:

$$p_{2,A\cap B} = \frac{nMOS_i}{MOS_A} \times \frac{MOS_A}{MOS_A + MOS_B}$$

$$p_{2,A\cap B} = \frac{nMOS_i}{MOS_A + MOS_B} \qquad (4)$$

where n is the number of PSUs to sample from the collapsed PSU stratum in the next scenario, which will equal 2 except for in-between scenarios, in which case n could equal 2, 3, or 4.

The unconditional probabilities can be used to calculate the conditional probabilities described as follows:

$$p_{k|A} = \frac{p_{k,A\cap B}}{p_{k-1,A}} \qquad (5)$$

where $p_{k,A\cap B}$ is the unconditional probability of the PSU in the collapsed stratum A in scenario k and $p_{k-1,A}$ is the unconditional probability of the PSU in the previous scenario (*k-1*).

5. Sampling PSUs to Ensure a Nested Design

To ensure a nested design for flexibility, PSU sampling needed to be done in stages, one for each scenario for both CRSS and CISS. CRSS only had 5 stages, but CISS had 7 stages since two additional scenarios were added. For Scenario 1, using the entire PSU frame, 2 PSUs were sampled per stratum with PPS to get a sample of 101 PSUs for CRSS and 49 PSUs for CISS. This can be referred to Stage 1. To ensure proper nesting of samples, Scenario 2 used the PSU sample from Scenario 1 as its PSU sampling frame. Then, using the conditional probabilities calculated, two PSUs per stratum were sampled for Scenario 2. Similarly, for Scenario 3, the sampling frame of PSUs was the PSU sample selected from Scenario 2, and the conditional probabilities were used to sample two PSUs per stratum for Scenario 3. The same was done for Scenarios 4 and 5.

After this sampling was completed, Scenarios 0 and 0.5 were added to CISS. These were obtained from Scenario 1 by sampling additional PSUs from the strata defined for Scenario 1. The PSUs in Scenario 1 were sampled using systematic sampling with probability proportional to size, and the sampling interval was halved to obtain the additional PSUs for Scenario 0. Scenario 0.5 was obtained by subsampling one of the additional PSUs added for Scenario 0.

6. Flexibility for Intermediate Scenario PSU Sample Sizes

The scenarios allowed for varying PSU sample sizes to an extent, but each scenario itself had a fixed number of sampled PSUs. To have a truly flexible design, a mechanism needed to be created to allow for PSU sample sizes in between the scenario sample sizes. This was done using a sort order of the PSUs and implementing the PSUs in their sort order. Three different methods for determining the sort order were explored. The random sort order was used for all CRSS scenarios and Scenarios 1-5 of CISS; the rank order of variance method was used for Scenarios 0 and 0.5 of CISS.

6.1 Random Sort Order

The first method of determining the sort order explored was by using a random sort order. The Scenario 5 sampled PSUs were first randomly sorted, and therefore would have sort order of 1-16 for both CRSS and CISS. Then, the Scenario 4 sampled PSUs that were not sampled in Scenario 5 were randomly sorted to have sort order of 17-24 for both CRSS and CISS. Likewise, the Scenario 3 sampled PSUs that were not sampled in Scenario 4 were randomly sorted to have sort order of 25-51 for CRSS and 25-32 for CISS. The same process was done for the remaining scenarios.

6.1.1 Advantages to a Random Sort Order

By randomly sorting the PSUs, the overall unconditional probabilities of selection are maintained. Also, no intentional biases with respect to the distribution of PSUs are introduced by the order.

6.1.2 Disadvantages to a Random Sort Order

Randomly sorting the PSUs could result in intermediary samples with unintentional and undesirable characteristics. For example, the first 4 PSUs sorted could all be from the same major stratum, and even from the same PSU stratum.

Therefore, if an in-between sample size was used, the first few PSUs could all be from the West urban major stratum, causing an apparent imbalance relative to the nation. This disadvantage could be avoided by randomizing the order of the major strata for the additional PSUs.

6.2 Rank Order of Variance of PSU Stratum

The next method of determining the sort order explored was using a rank order of a PSU stratum's contribution to the total, overall variance in each scenario. Variance of the Scenario 5 PSU strata estimated, and then ranked 1-8 for both CRSS and CISS. Then, one PSU was randomly sampled from the stratum ranked 1 and given order 1, one PSU was randomly sampled from the stratum ranked 2 and given order 2, and so on until one PSU was sampled from each stratum with the sort order matching the rank of the stratum. Then, using the same stratum ranks, the 2nd PSU from the stratum ranked 1 was given order 9; the 2nd PSU from the stratum ranked 2 was given order 10, etc., until all PSUs in Scenario 5 had an order. The same process was repeated for all scenarios and all PSUs.

6.2.1 Advantages to a Rank Order of Variance

By using the rank order on variance, the in-between scenario sample size is allocated to maximize the marginal increase in precision. This is because the PSU stratification is known and fixed, the sampling of PSUs within PSU strata is independent from PSU stratum to PSU stratum, and each PSU has a known non-zero probability of selection within each PSU stratum.

6.2.2 Disadvantages to a Rank Order of Variance

The sort order could still have the same undesirable characteristics as with a random sort, meaning that all urban strata could be sampled before rural strata based on the variance of the strata. However, this does not have as strong of an effect as randomly sorting PSUs because the two PSUs from the same stratum would never be sampled consecutively.

6.3 Rank Order of MOS of PSU Stratum

The third method explored was sorting by the rank order of the MOS of the PSU stratum in each scenario. The total MOS of the Scenario 5 PSU strata was ranked 1-8 for both CRSS and CISS. Then, 1 PSU was randomly sampled from the stratum ranked 1 and given order 1, 1 PSU was randomly sampled from the stratum ranked 2 and given order 2, etc. until 1 PSU was sampled from each stratum with the sort order matching the rank of the stratum. Then, using the same stratum ranks, the 2nd PSU from the stratum ranked 1 was given order 9; the 2nd PSU from the stratum ranked 2 was given order 10, etc., until all PSUs in Scenario 5 had an order. The same process was repeated for all scenarios and all PSUs.

6.3.1 Advantages to a Rank Order of MOS

By using the rank order of MOS, the in-between scenario sample size could potentially ensure the number of cases required for the second and third stages of sampling, as well as ensure no unintentional ordering occurred that could have occurred with a random sort. The former could have some advantages for CISS caseloads.

6.3.2 Disadvantages to a Rank Order of MOS

The sort order could still have characteristics similar to those found with a random sort, meaning that all urban strata could be sampled before rural strata based on the MOS of the strata. However, this does not have as strong of an effect as randomly sorting PSUs because the two PSUs from the same stratum would never be sampled one right after the other. Also, sorting by the rank order of MOS does not assist with minimizing variance, which is an objective of the redesign.

7. PSU Stratification for PSU Sample Sizes in Intermediary Scenarios

For PSU sample sizes in-between scenarios, PSU stratification is always determined by the smaller PSU sample size scenario. For example, if 60 PSUs for CRSS are to be implemented, this sample size is between scenarios 2 and 3. The weights for these PSUs are to be calculated as the inverse of the conditional probabilities from equation 3 where the stratum of the PSU is that of scenario 3. This would ensure each noncertainty stratum has at least 2 PSUs sampled, which is required for unbiased variance estimation.

8. Conclusion

A flexible, scalable PSU sample design was developed for both the CRSS and CISS to make them adaptable for use with different levels of resources. Depending on the PSU sample size, the PSU sample design could have between two and four PSUs sampled per PSU stratum. The PSUs sampled have known probabilities of selection. A mechanism using ordering allows any sample size between 16 and 101 for CRSS and between 16 and 96 for CISS to be implemented. The nested design allows for a changing budget, where PSUs can be added or removed in the future.

Acknowledgements

I would like to thank Fan Zhang, Chou-Lin Chen, Tina Morgan, and Donna Glassbrenner of NHTSA and Rui Jiao, Yumiko Sugawara, and Fran Bents of Westat.

References

Krenzke, T., and Haung, W.-C. (2009). Revisiting Nested Stratification of Primary Sampling Units. *Federal Committee on Survey Methodology Conference Proceedings*, https://fcsm.sites.usa.gov/files/2014/05/2009FCSM_Krenzke_IX-C.pdf.

Moriarity, C., and Parsons, V. (2013). Expanding the Number of Primary Sampling Units for the National Health Interview Survey. *American Statistical Association Joint Statistical Meetings*.