B. COLLECTIONS OF INFORMATION EMPLOYING STATISTICAL METHODS

1. Describe the potential respondent universe and any sampling or other respondent selection methods to be used.

The CISS universe, or sampling frame, is the set of police-reported motor vehicle crashes on a traffic way involving a passenger vehicle and in which a passenger vehicle is towed from the scene resulting a police accident report (PAR). The estimated CISS population size is about 1.9 million a year. CISS samples from this basic frame through a stratified multi-stage probability sampling scheme as follows:

Divide the country into geographic units called Primary Sampling Units (PSUs). A PSU is a county or group of counties. PSUs were formed as groups of adjacent counties with end-to-end distance no more than 65 miles for urban area and 130 miles for rural area to ensure efficient data collection. PSUs were also formed in such a way that with 90% of chance there are at least 5 fatal crashes involving a passenger vehicle in a given year. PSU formation respects region and urbanicity boundary. Some outlying areas of Alaska and small islands of Hawaii were excluded. There are total 1,784 PSUs in the PSU frame.

The PSU frame is then stratified into 8 primary PSU strata by two variables – region (Northeast, West, South, and Midwest) and urbanicity (urban and rural). Within each primary stratum, PSUs are further stratified by other secondary stratification variables such as Road Type, Total Crash, and Total Vehicle Mile Traveled. A composite measure of size (MOS) value was assigned to each PSU in the frame. CISS PSU MOS is a linear combination of estimated seven high interest crash counts (e.g. fatal crash, incapacitated injury crash, late model year vehicle crash etc.) of the PSU. PSUs with similar characteristics were grouped into secondary strata with approximately equal MOS sizes and minimum within stratum variances. As the result, total 24 PSU strata are formed.

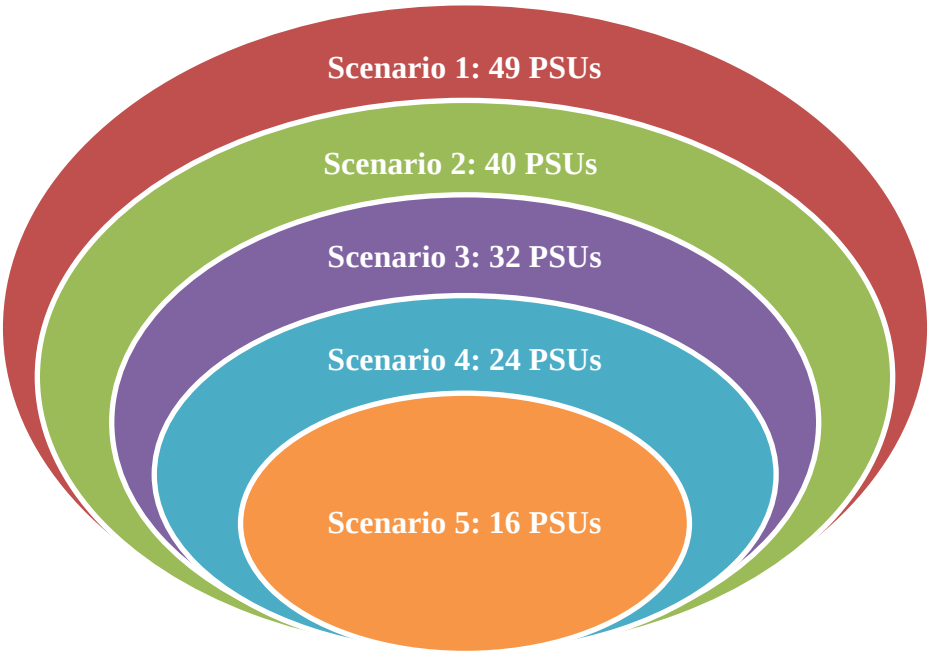
A probability proportional to size (PPS) sample of two PSUs per stratum is selected systematically from each of the 24 PSU strata. The resulting sample with 48 non-certainty PSUs and one certainty PSU serves as the first phase PSU sample.

We anticipate no more than 49 PSUs will actually be used for CISS data collection due to budget constraints. The actual CISS PSU sample size is smaller than 49. To cope with the uncertainty of future budget level, a sequence of nested PSU sub-samples were selected from the 49 PSU first phase sample. For example, the second phase PSU sample has 40 PSUs and 20 PSU strata. To select the second phase PSU sample, four first phase PSU strata were collapsed with other strata to form total 20-second phase

strata. Then two PSUs were selected from the collapsed first phase PSU sample in each of the 20-second phase strata. Total five phases of PSU sample were selected – each is a sub-sample of previous phase PSU sample with sample size range from 16 to 48 plus one certainty PSU (see figure and table below). The PSU collapsing order is predetermined. Each sample can be used as PSU sample and the resulting PSU selection probability is PPS. This approach produces flexible PSU sample and minimize the impact to existing PSU sample when sample size changes. Typically, non-certainty PSU selection probability is:

$$\pi_{hi} = \frac{2\,MOS_{hi}}{\sum_{i} MOS_{hi}}$$

Here h is PSU stratum, i is PSU, MOS_{hi} is PSU MOS, summation is over the non-certainty PSUs in the stratum. The CISS PSU sample size can be expanded beyond 49 if needed.



Scenario	Number of PSU Strata	Number of Certainty PSU	Total Number of Sampled PSUs	Sampled PSU per Stratum
1	24	1	49	2
2	20	0	40	2
3	16	0	32	2
4	12	0	24	2
5	8	0	16	2

- Within each PSU, PARs are stratified by the police jurisdictions (PJ) where PARs are available and PJs become the second stage sampling units. A composite MOS is assigned to each PJ in the selected PSUs. PJ MOS is a summation of PAR domain crash count percentages in the PSU weighted by the predetermined PAR domain sampling rate. So PJs with larger desirable crash composition are selected with larger probabilities. PJs are then stratified into two PJ strata by their MOS (largest 50% vs the rest) in addition to certainty PJs. A PJ sample is then selected from each PJ stratum using sequential Poisson sampling.

  Sequential Poisson sampling method (see Ohlsson, Esbjörn (1998): Sequential Poisson Sampling, Journal of Official Statistics, Vol.14, No.2, pp. 149–162) produces an approximate PPS sample, handle the frame changes and minimize the changes to the existing sample at the same time.

  Sequential Poisson sampling method was applied to the PJ sample selection for each of non-certainty PJ strata (large MOS or small MOS stratum) within the sampled PSUi, as following:

  - Generate a permanent uniform random number r_{ij} $U(0,1)$ for each PJ j in the PJ frame.
  - Identify certainty PJs by the condition:

    $$\frac{m_i * MOS_{ij}}{\sum_{j=1}^{M_i} MOS_{ij}} \geq 1$$

    Here m_i is the PJ sample size and M_i is the PJ frame size for a PJ stratum within PSU i. MOS_{ij} is the PJ MOS. The identified certainty PJs are set aside. And this process is repeated to the remaining PJs based on the reduced PJ sample size until there is no more certainty PJs. Let the total number of certainty PJs be m_c.
  - For the remaining $M_i - m_c$ non-certainty PJs in the frame, divide their permanent random number by the MOS to obtain the transformed random number: r_{ij}/MOS_{ij}. Then, sort the transformed random number from the smallest to the largest as following:

    $$\frac{r_{i1}}{MOS_{i1}} \leq \frac{r_{i2}}{MOS_{i2}} \leq \cdots \leq \frac{r_{i(M_i - m_c)}}{MOS_{i(M_i - m_c)}}$$
  - Thus, the m_c certainty PJs plus the first $m_i - m_c$ non-certainty PJs on the above list are the PJ sample for a PJ stratum within PSU i.

  Sequential Poisson sampling is approximately PPS. The PJ selection probability is:

$$\pi_{j \vee hi} = \frac{m_{hi} MOS_{hij}}{\sum_{j} MOS_{hij}}$$

Here j is for PJ, m_{hi} is the PJ sample size for PSU i, MOS_{hij} is the PJ MOS. The summation is over all non-certainty PJs in the selected PSU.

- PARs are selected on a weekly basis. In other words, the crashes over the year are stratified by weeks. Each week in each selected PSU, technicians visit the selected PJs and list PARs recorded at that jurisdiction since the last visit. During the PAR listing process, PARs are labeled into 10 analysis domains. For some large PJs with large number of PARs, a systematic sample of PARs is listed. If one of every L PARs is sub-listed in PJ j, PSU i, the sub-listing probability for all sub-listed PARs are:

  $$\pi_{l \vee hij} = \frac{1}{L}$$

  After PARs in all sampled PJs are listed, all listed PARs in the same PSU are pooled together. A PAR MOS is assigned to every listed PAR. PAR MOS is the product of a PAR domain MOS factor, the PJ design weight, and the sub-listing factor (the inverse of $\pi_{l \vee ij}$). PAR domain MOS factor is determined by simulation to ensure the predetermined PAR domain sample sizes can be achieved. PARs in rare domain and/or smaller PJ receive larger MOS.

  A PAR sample of average 1.75 crashes per data collector in each PSU every week is selected using sequential Poisson sampling. Sequential Poisson sampling produces a scalable sample so it allows us to replace the non-responding cases (the cases with key vehicle information missing) and to better handle workload changes. Under sequential Poisson sample selection, the PAR selection probability is:

  $$\pi_{k \vee hijl} = \frac{k_{hi} MOS_{hijk}}{\sum_{k} MOS_{hijk}}$$

  Here k_{hi} is the weekly PAR sample size for PSU i, MOS_{hijk} is PAR MOS. The summation is over all listed non-certainty PARs of the week in the PSU.

- The overall selection probability is:

  $$\pi_{hijlk} = \pi_{hi} * \pi_{j \vee hi} * \pi_{l \vee hij} * \pi_{k \vee hijl}$$

  The design weight is the inverse of π_{hijlk}.

- PSU, PJ and PAR sample sizes are estimated using optimization by minimizing variance subject to cost assuming three-stage simple random sampling without replacement.

  The optimization model consists of the objective function, cost constraint, and variance constrains as following.

$$Minimize: \sum_{g=1}^{G} V_{CISS}(\bar{\bar{\bar{y}}}_g) = \sum_{g=1}^{G} \left\{ \frac{S_{1,g}^2}{n}\left(1-\frac{n}{N}\right) + \frac{S_{2,g}^2}{nm}\left(1-\frac{m}{M}\right) + \frac{S_{3,g}^2}{nmk}\left(1-\frac{k}{K}\right) \right\}$$

$$Subject\ ¿: C = C_0 + n C_1 + nm C_2 + nmk C_3,$$

$$V_{CISS}(\bar{\bar{\bar{y}}}_g) = \frac{S_{1,g}^2}{n}\left(1-\frac{n}{N}\right) + \frac{S_{2,g}^2}{nm}\left(1-\frac{m}{M}\right) + \frac{S_{3,g}^2}{nmk}\left(1-\frac{k}{K}\right)$$

$$\leq V_{CDS}(\bar{\bar{\bar{y}}}_g), for\ g = 1, \cdots, G.$$

$mk \geq l$.

g: Subscript of the identified key estimate, $g = 1, \cdots G$. Here $G = 7$.

$\bar{\bar{\bar{y}}}_g$: Identified key proportion estimate.

n, m, k: Optimal sample sizes of PSUs, PJs, and cases (PARs) to be determined.

N: Population size of PSUs

M: Average population size of PJs.

K: Average population size of PARs

$V(\bar{\bar{\bar{y}}}_g)$: Variance of the identified key estimate $\bar{\bar{\bar{y}}}_g$.

$S_{1,g}^2, S_{2,g}^2, S_{3,g}^2$: Variance component at PSU-, PJ-, and case-level.

C, C_0, C_1, C_2, C_3: Total, fixed, PSU-, PJ-, and crash-level cost coefficients.

$V_{CDS}(\bar{\bar{\bar{y}}}_g)$: Variance of the identified key estimate $\bar{\bar{\bar{y}}}_g$ in the current system (NASS CDS).

l: known case load.

- Under the CDS cost components and assume two (2) technicians per PSU, it was determined the optimum sample allocation is about 8 PJs per PSU, and 200 cases per PSU. Under the current CISS budget, NHTSA expects to collect data from thirty-two (32) PSUs.

- Standard errors for seven key estimates under current CDS were used as constraints in the above optimization model to ensure the corresponding degree of accuracy under CISS will be at least as good as CDS.

- Once crashes are selected, a detailed investigations is conducted on the selected crashes.

NHTSA collects crash counts from the non-sampled PJs in the selected PSUs to improve the accuracy of the national estimates.

Expected completion rates for the additional investigation stages are scene inspection 98%; vehicle inspection 85%; occupant interview 83%; and, occupant injury 88%. If the key vehicle is missing, a replacement PAR is selected for investigation from the listed PARs sorted by

sequential Poisson sampling method. This dramatically increases the effective sample size and data quality.

After design weights are calculated, the weights need to be adjusted for the following reasons:

- Refusal/non-respondent adjustment at all 3 stages;
- Frame coverage bias correction;
- Large weight trimming;

Calibration technique will be used as the adjustment method. The potential auxiliary information to be used for calibration includes Census population counts and PSU level total crash counts.

The calibration adjustment method that handle all the above have been implemented in SUDAAN 11 WTADJX procedure. SUDAAN WTADJX procedure will be used to create the final analysis weights.

Some key item missing values will be imputed. Several imputation methods will be considered and used for imputation, depending on the missing variable and available information. The imputation methods include but not restricted to: logical imputation, regression imputation, and hot deck imputation.

Since the resulting PSU sampling rate is quite low, we expect the PSU sample selection can be approximated treated as with-replacement sample selection. The standard specialized software such as SAS SURVEY procedures and SUDAAN procedures can be used for CISS data analysis.

2. Describe collection of information procedures.

Once a crash has been selected for investigation, the CISS team initiates several activities. Researchers locate, visit, measure, and photograph the crash scene; locate, inspect, and photograph all involved vehicles; conduct a telephone or personal interview with each involved person or surrogate; and, record injury information from hospitals or emergency rooms for all injured victims. During each activity, the researchers record information on the crash, vehicle, and occupant forms as appropriate.

3. Describe methods to maximize response rates and to deal with issues of non-response.

The Crash Investigation Sampling System (CISS) has a three-stage sample design. The first stage sampling units are counties or groups of counties. A PSU becomes a non-responding PSU only if all selected police jurisdictions (PJs) within the PSU are non-responding PJs. In 2017, all sampled PSUs responded and we expect so in the future years. In the CISS, PJ samples are selected using sequential Poisson sampling method. The whole PJ frame can be used as replacement sample. Therefore, a PSU becomes non-responding PSU only if all PJs in the frame are non-responding PJs. Therefore, by design, it is unlikely there will be any non-responding PSUs in the CISS.

The second stage sampling units of CISS are PJs. A sampled PJ becomes non-responding PJ if it refuses to cooperate. To improve PJ cooperation rate, NHTSA plan to visit each selected PJ and meet with local law enforcement officers to gain cooperation. In 2017, 95% sampled PJs in the CISS responded and we expect the similar response rate in the future years.

At the third stage, first all police accident reports (PARs) in the selected PJs are listed, then a sample of PARs is selected by Pareto sampling. If the key vehicle is not available for inspection, we replace it with another case. In 2017, 294 out of 2069 selected cases were replaced. This is a fourteen (14) percent replacement rate.

The item response rate of 2017 CISS varies. For example, scene inspection variables have item response rates around 98%; towed vehicle inspection 88%; occupant interview 66%; and, occupant injury 76%. CISS quality control system will be designed to produce the most accurate, reliable, and complete database possible within the limits of available resources. All data will be automated and edited by a complex algorithm which checks for inconsistencies and questionable items. A sample of all crashes will be given a thorough review by an experienced researcher at a Data Quality Control Zone Center. Zone Center personnel will visit each PSU regularly to observe the team's investigation activities and to discuss systematic problems revealed in edit and Zone Center reviews of the team's cases.

Since the interview is vital to a complete case, CISS teams will make special efforts to complete an interview when at all possible. Occupants will be contacted by telephone. CISS researchers will call at varying hours (often in evenings or on weekends) until they have located the person sought. When the person is unavailable, other passengers or witnesses are contacted. If the person sought cannot be located by telephone, researchers use personal visits

or mail questionnaires. Each CISS researcher will be given special training in interviewing. This increases the possibility that persons will cooperate once they have been located and contacted. As a result of these procedures used in our legacy program (NASS), it's anticipated that CISS teams will complete more than three-quarters of all occupant interviews. Proposed interview forms to be used for CISS are displayed in Attachment 7.

As a final check on CISS data, approximately 5% of those interviewed will be recontacted by Zone Center personnel to establish that they had in fact been interviewed and to verify some of their responses. This type of interview takes approximately 5 minutes.

4. Describe any tests of procedures or methods to be undertaken.

NHTSA will test new data collection procedures for six (6) months. The test will include: gathering police crash reports and identifying qualified crashes to investigate, conducting the interviews to assess interview questions and procedures, entering data from scenes and vehicles with electronic devices, analyzing data, and monitoring for quality control.

The attached spreadsheet (Attachment 3) shows the list of data elements collected from the detailed investigation of the selected crashes. The electronic forms and protocols have been developed to collect this information on tablet computers.

5. Provide the name and telephone number of individuals consulted on statistical aspects of the design and the name of the agency unit, contractor(s), grantee(s), or other person(s) who will actually collect and/or analyze the information for the agency.

Ms. Chou-Lin Chen, National Center for Statistics and Analysis, NHTSA, 202-366-1048 is responsible for CISS survey design.

NHTSA has decided to undertake a basic redesign of the National Automotive Sampling System that will attempt to meet new and diverse requirements through expanding its scope and making it more responsive to changing needs. Accordingly, NHTSA contracted with Westat (contract DTNH22-12-F-00389) to lead the survey modernization effort, but also participated jointly with Westat in developing the new Crash Investigation Sampling System (CISS). The CISS

contractors are Calspan Corporation (contract DTNH22-14-D-00363) and KLD Associates, Inc. (contract DTNH22-14-D-00366).