U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

NHTSA

DOT HS 812 801

September 2019

Crash Investigation Sampling System: Design Overview, Analytic Guidance, and FAQs

Suggested APA Format Citation:

Zhang, F., Subramanian, R., Chen, C.-L., & Young Noh, E. Y. (2019, September). *Crash Investigation Sampling System: Design overview, analytic guidance, and FAQs* (Report No. DOT HS 812 801). Washington, DC: National Highway Traffic Safety Administration.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.	
DOT HS 812 801			
4. Title and Subtitle Crash Investigation Sampling System: Design Overview, Analytic Guidance, and FAQs		5. Report Date September 2019	
		6. Performing Organization Code NSA-210	
7. Author(s) Fan Zhang, Rajesh Subramanian, Chou-Lin Chen, Eun Young Noh		8. Performing Organization Report No.	
9. Performing Organization Name Mathematical Analysis Division, National Center for Statistics and Analysis National Highway Traffic Safety Administration 1200 New Jersey Avenue SE Washington, DC 20590		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No.	
12. Sponsoring Agency Name and Address Mathematical Analysis Division, National Center for Statistics and Analysis National Highway Traffic Safety Administration 1200 New Jersey Avenue SE. Washington, DC 20590		13. Type of Report and Period Covered NHTSA Technical Report	
		14. Sponsoring Agency Code	
15. Supplementary Notes The authors would like to thank Phillip Kott for his consultation.			
Abstract This document describes the Crash Investigation Sampling System (CISS) sample design and weighting procedures and explains some basic concepts about estimation based on complex survey data. In addition, it provides examples and discusses issues of CISS data analysis.			
17. Key Words NHTSA, CISS, CDS, NASS, sample design, complex survey data analysis, analytic guidance.		18. Distribution Statement This document is available from the National Technical Information Service www.ntis.gov.	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 29	22. Price

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Acronyms

- AUM – analytical user's manual
- CDS – Crashworthiness Data System
- CISS – Crash Investigation Sampling System – a replacement of CDS
- CRSS – Crash Report Sampling System – a replacement of GES
- FARS – Fatality Analysis Reporting System
- GES – General Estimates System
- JK – Jackknife
- MOS – Measure of Size
- NASS – National Automotive Sampling System
- PAR – police crash/accident report
- PJ – police jurisdiction
- PSU – Primary Sampling Unit
- SSU – Secondary Sampling Unit
- TSU – Tertiary Sampling Unit

TABLE OF CONTENTS

1. Introduction

The National Highway Traffic Safety Administration developed and implemented the National Automotive Sampling System in the 1970s to make estimates of the motor vehicle crash experience in the United States. In 1988 NHTSA split the NASS into two surveys, the General Estimates System and the Crashworthiness Data System. Since then the same data collection sites have been used for GES and CDS data collection. Given the shifts in population and the vehicle fleet, and the changing analytic needs of the safety community, Congress authorized NHTSA to modernize its crash data collection system.

NHTSA implemented two new annual surveys, the Crash Report Sampling System - which replaced the GES, and the Crash Investigation Sampling System - which replaced the CDS.

This document first provides an overview of the CISS sample design (Chapter 2) and weighting procedure (Chapter 3). The sample design and weighting procedures determine the design option associated with the variance estimation method to be used in our examples and need to be accounted if users choose to use other alternative analysis methods.

Chapter 4 describes some basic concepts on the analysis of complex survey data which justifies the practice of using finite-population point estimates and design variance estimates when making inferences about model parameters.

Chapter 5 provides examples of making estimates using CISS and CDS data and discusses issues related to CISS data analysis. Finally, Chapter 6 catalogs frequently asked questions and answers on sampling and estimation of CDS/CISS.

While this document provides a broad overview of the design of CISS, a supplemental NHTSA technical report, *Crash Investigation Sampling System: Sample Design and Weighting* (Zhang et al., in press) to be published by NHTSA describes the CISS sample design and weighting procedures in greater detail.

2. The CISS Sample Design

CISS was designed independent of other NHTSA surveys. The target population for the CISS is all police-reported motor vehicle crashes on a traffic way, each involving a passenger vehicle[1] and in which a passenger vehicle is towed from the scene for any reason. This definition is slightly different from the CDS, which required that a vehicle be towed due to damage. This change was made because sometimes it was difficult to determine why a vehicle was towed.

Because a direct collection of crashes in the nation is infeasible, the CISS crash sample is selected in multiple stages to produce a nationally representative probability sample.

At the first stage, 3,117 counties in the United States were grouped into 1,784 Primary Sampling Units. A PSU in the CISS is either a county or a group of counties. U.S. territories, some remote counties in Alaska, and small islands of Hawaii were excluded. PSUs have been formed in such a way that there is a 90 percent chance to have at least 5 fatal crashes every year inside each PSU and the end-to-end distance of a PSU was 65 miles for an urban area and 130 miles for a rural area.

The 1,784 PSUs were stratified into 24 strata by the four Census regions, urban/rural, total highway/primary/secondary road miles, and total expected number of crashes. Each of the 1,784 PSUs in the frame was assigned a measure of size (MOS) equal to the combination of its estimated seven types (defined by injury severity and vehicle model year) of crash counts.

From each of the 24 PSU strata, 2 PSUs were selected by a probability proportional-to-size (PPS) sampling method. In addition, one large PSU was selected with certainty. This resulted in a total of 49 PSUs. Then a sequence of PSU sub-samples was selected from the 49 PSUs with decreasing sample sizes. In this process the PSU strata were collapsed when necessary. This process produced a sequence of nested PSU samples. These nested PSU samples allow NHTSA to change the PSU sample size without reselecting the sample. The final PSU sample is the result of multiphase sampling and the PSU sample selected in such a way remained generally PPS.

For the 2017 CISS, for example, 24 PSUs were selected from 12 PSU strata (2 PSUs selected per stratum) using the above process. Consequently, the PSU sampling rate was very low in each stratum. Because of the reduction of PSU sample size (from 49 to 24), no PSU was selected with certainty. All sampled PSUs cooperated with NHTSA's data collection request.

The Secondary Sampling Units are police jurisdictions. Within each selected PSU, PJs were stratified into three PJ strata by their estimated measure of size - a combination of crash counts in six categories of interest. The Pareto sampling method (Rosén, 1997) was used to select PJ samples from each PJ stratum. The Pareto sampling method produces overlapping samples when a new sample is reselected. This reduces the changes to the existing PJ sample when a new PJ sample needs to be selected because of PJ frame (the collection of all PJs in the selected PSU) changes. The PJ inclusion probability under the Pareto sampling is approximately PPS (Rosén, 1997). In 2017 CISS, across the 24 sampled PSUs, a total of 182 PJs were selected and 168 PJs

[1] CISS-applicable vehicles are the same as CDS-applicable vehicles: passenger cars, light trucks, vans, and sport utility vehicles with gross vehicle weight rating (GVWR) less than 10,000 lbs.

cooperated. Weight adjustments were made to mitigate the potential bias caused by the 14 non-responding PJs.

The Tertiary Sampling Units are PARs. Every week, the CISS data collectors receive PARs accumulated since the last sample selection from all the selected PJs in the same sampled PSU. All new PARs are grouped into ten PAR domains (see Table 1). These ten PAR domains are formed based on the results of NHTSA's internal and public data needs. The PAR domains are used to oversample the following important analysis domains to ensure enough cases are selected into the sample:

- Crashes involving occupant killed;
- Crashes involving occupant injured or possibly injured in a recent model year passenger vehicle (vehicle no more than 4 year old); and
- Crashes involving occupant severely injured in a passenger vehicle.

An MOS is then assigned to each PAR. This MOS is PSU-, PJ-, and PAR-domain-specific. It is determined to ensure the desired sample allocation defined in Table 1 can be achieved.

All the PARs in the same PSU are pooled together, and a PAR sample is selected using a Pareto sampling method. As in CDS, CISS PAR sample selection is conducted weekly.

After the initial PAR sample has been selected, if a vehicle that defines the selected PAR's domain is unavailable for data collection, the original PAR sample size is augmented and the PAR sample is reselected using a Pareto sampling method so a replacement PAR can be included to replace the non-responding PAR. In 2017 there were 288 PARs added to the original sample to replace the non-responding PARs. This resulted a total of 2,331 selected cases. After excluding the non-responding cases and the 8 out-of-scope cases, a total of 2,035 cases were kept in the final analysis file. This augment increases the selected sample size while keeping the respondent (i.e., investigated) sample size approximately equal to the initial target PAR sample size within each PSU.

For each selected PAR, CISS technicians collects information about the crash, the vehicles involved in the crash, and the occupants involved in the crash. Trained crash investigators obtain data from crash sites by studying crash evidence such as skid marks, fluid spills, broken glass, and bent guard rails. They locate the vehicles involved, photograph them, measure the crash damage, and identify interior locations that were struck by the occupants. The researchers also interview crash victims and review their medical records to determine the nature and severity of injuries.

Because of the low PSU sampling rates, PSU sample can be approximately viewed as selected with-replacement. This simplifies the variance estimation.

For more details about CISS sample design, see the upcoming NHTSA technical report, *Crash Investigation Sampling System: Sample Design and Estimation* (Zhang et al., in press).

Table 1: CISS PAR Domains, Crash Sample Size Allocation, and Population Estimates

CISS Analysis Domains	Description	Target Percent of Sample Allocation	Estimated Population	Population Percent
1	At least one occupant of towed passenger vehicle is killed	5%	9,576	0.51%
2	Crashes not in Stratum 1 involving: • A recent model year passenger vehicle in which at least one occupant is incapacitated	10%	17,304	0.93%
3	Crashes not in Stratum 1 or 2 involving: • A recent model year passenger vehicle in which at least one occupant is non-incapacitated, possibly injured or injured but severity is unknown.	20%	162,037	8.71%
4	Crashes not in Stratum 1-3 involving: • A recent model year passenger vehicle in which all occupants are not injured	15%	325,332	17.48%
5	Crashes not in Stratum 1-4 involving: • A mid-model year passenger vehicle in which at least one occupant is incapacitated	6%	23,739	1.28%
6	Crashes not in Stratum 1-5 involving: • A mid-model year passenger vehicle in which at least one occupant is non-incapacitated, possibly injured or injured but severity is unknown	12%	210,407	11.31%
7	Crashes not in Stratum 1-6 involving: • A mid-model year passenger vehicle in which all occupants are not injured	10%	418,702	22.51%
8	Crashes not in Stratum 1-7 involving: • An older model year passenger vehicle in which at least one occupant is incapacitated	6%	28,690	1.54%
9	Crashes not in Stratum 1-8 involving: • An older model year passenger vehicle in which at least one occupant is non-incapacitated, possibly injured or injured but severity is unknown.	10%	220,815	11.87%
10	Crashes not in Stratum 1-9 involving: • An older model year passenger vehicle in which all occupants are not injured	6%	443,151	23.83%
Total		100%	1,859,752	100%

Source: The population estimates were made using 2011 CDS data.

Note: This table uses the following definitions:

- Recent Model Year (or Late Model Year) Vehicles: Vehicles that are <= 4 years old.
- Mid-model Year Vehicles: 5- to 9-year-old vehicles
- Older Model Year Vehicles: Vehicles that are 10 years old or older

3. CISS Weighting Procedures

The CISS sample is the result of probability sampling featuring stratification, clustering, and selection with unequal probabilities. Because of these complex design features, the CISS sample is not a simple random sample and users need to use proper weights to produce estimates reasonably robust against selection biases. The 2017 CISS weights were created with the following steps:

- Calculate the base weights (the inverse of selection probabilities) at all three stages (PSU, PJ, and PAR).
- Adjust the base weights for PJ and PAR non-response[2] to correct potential nonresponse bias.
- Calibrate the PJ and the PAR weights using the PSU level total PAR stratum PAR counts to further correct for potential Pareto weighting error, nonresponse and coverage biases, and increase the precision of the estimates.
- Calibrate PSU weights to capture population shift.
- Truncate the large case weights. Case weights larger than 3 percent of the PAR domain weight total are truncated to 3 percent of the PAR domain weight total and the excessive weights are redistributed to other untruncated weights in the same PAR domain (a form of additional calibration).
- Adjusted Jackknife replicate weights are created. Jackknife variance estimation method assumes the PSU sample has been selected with-replacement. These adjusted Jackknife replicate weights capture the impacts of weight adjustments in the variance estimation.

The final weight variable for the CISS estimation is CASEWGT. The Jackknife replicate weights are JKWGT1 - JKWGT24 for 2017.

See Crash Investigation Sampling System: Sample Design and Weighting, (Zhang et al., in press) for more detailed information on the CISS weighting procedure.

[2] A PAR is non-responding if the vehicle that defines the PAR's domain is unavailable for inspection. Non-responding PJs are the PJs that refused to cooperate.

4. Basic Concepts of Complex Survey Data Analysis

In this chapter, we introduce some basic concepts of complex survey data analysis which explains why we can estimate a model parameter by estimating a related finite-population parameter and use the sampling variance with respect to the sample selection to approximate the total variance. This discussion justifies the approach we use in all our examples.

4.1 Model Parameter Estimation

In standard statistical theory, we often assume that the data generated by nature or by a laboratory experiment follows a stochastic model. The model parameter that indexes the underlining model is of interest and needs to be estimated. For example, consider fatal indicators $\{y_1, y_2, \dots, y_N\}$:

$$y_k = \begin{cases} 1, & fatal\ crash \\ 0, & nonfatal\ crash \end{cases}, \quad k = 1, 2, \dots N$$

observed from the N CISS crashes reported in the year 2017. One may view these observations as outcomes of independent and identical Bernoulli trials indexed by model parameter θ:

$$y_k \sim Bern(\theta), \qquad k = 1, 2, \dots N$$

And use the maximum likelihood estimator:

$$\hat{\theta}_N = \frac{1}{N}\sum_{k=1}^{N} y_k$$

to estimate the model parameter θ. If this model is correct, $\hat{\theta}_N$ is unbiased with respect to the model for θ:

$$E_{Bern}(\hat{\theta}_N) = \frac{1}{N}\sum_{k=1}^{N} E_{Bern}(y_k) = \theta$$

with variance:

$$Var_{Bern}(\hat{\theta}_N) = \frac{1}{N^2}\sum_{k=1}^{N} Var_{Bern}(y_k) = \frac{\theta(1-\theta)}{N} = O(N^{-1}).$$

Here E_{Bern} and Var_{Bern} are the expectation and variance with respect to model $Bern(\theta)$. Notice when N is very large, the model variance $Var_{Bern}(\hat{\theta}_N)$ becomes very small.

4.2 Finite Population Parameter Estimation

In the previous section, the model parameter θ is estimated by:

$$\hat{\theta}_N = \frac{1}{N}\sum_{k=1}^{N} y_k.$$

However, the quantity $\hat{\theta}_N = \sum_{k=1}^{N} y_k/N$ itself is also of interest because it gives a snapshot of the nation's fatal crash proportion in 2017. Similar statistics include N (2017 total number of CISS crashes) and $\sum_{k=1}^{N} y_k$ (2017 total number of fatal CISS crashes) etc. In other words, in addition to model parameters, we may also be interested in the functions of a set of realized (fixed) values. For example, the collection of all realized 2017 CISS crashes $U = \{u_1, u_2, \dots, u_N\}$ can be viewed as a finite population. The functions of the attributes of the finite population, such as $\hat{\theta}_N$, N, and $\sum_{k=1}^{N} y_k$ are called finite population parameters.

Unfortunately, it is often cost-prohibitive to observe all the units in the finite population. Instead, a probability sample is selected and observed to estimate finite-population parameters.

A probability sample s is a subset of the finite population U selected under a probability sampling design. The key role of the probability sampling design is to define a probability space on U so we can use the sample s to estimate and make inferences about the finite population parameters. Chapters 2 and 3 briefly described how a probability sample of crashes was selected from a finite population of crashes for CISS data collection and how the final CISS weights were calculated.

It should be noted that for various reasons, it is inevitable to use design features such as stratification, clustering, and unequal selection probabilities to select the probability sample. For example, cluster sampling was used because it is not cost-efficient to obtain all crashes in the United States in order to directly select a one-stage crash sample. Crashes in important analysis domains were assigned larger selection probabilities to ensure enough sample sizes for analysis. Stratification was used at all stages to reduce the sampling variance and to produce more balanced sample (by assuring, for example, that sampled PSUs are found in all regions and urbanicities of the United States). These design features might induce a stochastic dependence among the resulting observations and alter the original distribution. As a result, the final sample is not a simple random sample, and the sampled observations may no longer follow the same model as the population from which they were drawn.

Under a probability sampling design, every unit u_k in the finite population $U = \{u_1, u_2, \dots, u_N\}$ has a positive probability π_k of being selected into the sample s. Assume sample $s = \{u_1, u_2, \dots, u_n\}$ has fixed sample size $n \leq N$ and define the selection indicator as:

$$I_k = \begin{cases} 1, & if\ u_k\ is\ selected\ into\ s \\ 0, & otherwise \end{cases} \qquad (k = 1,2, \dots, N)$$

The inverse of the inclusion probability $w_k = 1/\pi_k$ can be used to construct design-based point estimators of finite population parameters (i.e., they are unbiased or nearly unbiased under the

probability-sampling design). For example, let the fatal indicator y_k be an attribute observed from crash u_k, then

$$\hat{\theta}_n = \frac{1}{N}\sum_{u_k \in s} w_k y_k$$

is design unbiased for the 2017 fatality proportion: $\hat{\theta}_N = \sum_{k=1}^{N} y_k / N$:

$$E_D(\hat{\theta}_n) = E_D\left(\frac{1}{N}\sum_{u_k \in s} w_k y_k\right) = E_D\left(\frac{1}{N}\sum_{k=1}^{N} w_k I_k y_k\right) = \frac{1}{N}\sum_{k=1}^{N} y_k = \hat{\theta}_N$$

Here the expectation E_D is with respect to the probability space defined by the sampling design. The sampling/design variance of $\hat{\theta}_n$, $Var_D(\hat{\theta}_n)$, is the variance of estimator $\hat{\theta}_n$ under repeated probability sampling. $Var_D(\hat{\theta}_n)$ depends on both the estimator $\hat{\theta}_n$ and the sample design. It should be noted that the point estimator $\hat{\theta}_n$ is design unbiased for the finite population parameter $\hat{\theta}_N$ regardless of whether the model assumed to generate the finite population is true or not.

4.3 Two-Step Sampling Procedure

Combining the concepts in the two previous sections, survey data can be viewed as the result of the following two-step sampling procedure (Hartley & Sielken, 1975):

- Step 1: A finite population U of size N is generated by an infinite super-population model ξ.
- Step 2: A probability sample s of size $n \leq N$ is selected from the finite population U.

That is:

$$Model\ \xi \xrightarrow{Generation} U = \{u_1, u_2, \ldots, u_N\} \xrightarrow{Selection} s = \{u_1, u_2, \ldots, u_n\}$$

Under this two-step sampling view, the design unbiased point estimator is not only an unbiased estimator of the finite population parameter $\hat{\theta}_N$ under the probability based design, but also an unbiased estimator of the super-population model parameter θ if the (assumed) model is correct:

$$E_{\xi D}(\hat{\theta}_n) = E_\xi[E_D(\hat{\theta}_n)] = E_\xi[\hat{\theta}_N] = \theta$$

Here the expectation $E_{\xi D}$ is with respect to the two-step sampling process: the data generation by the model and the sample selection by the sample design. The total variance of a design unbiased point estimator $\hat{\theta}_n$ under this two-step sampling view can be decomposed as:

$$Var_{\xi D}(\hat{\theta}_n) = E_\xi[Var_D(\hat{\theta}_n)] + Var_\xi[E_D(\hat{\theta}_n)]$$

Since $E_D(\hat{\theta}_n) = \hat{\theta}_N$ and $Var_\xi(\hat{\theta}_N) = O(N^{-1})$, therefore $Var_\xi[E_D(\hat{\theta}_n)] = Var_\xi[\hat{\theta}_N] = O(N^{-1})$. So, when the finite population size N is large, the second term on the right is negligible. Therefore, if $\widehat{var}_D(\hat{\theta}_n)$ is a design unbiased estimator of $Var_D(\hat{\theta}_n)$, then it can also serve as an approximate estimator of the total variance when N is large:

$$\widehat{var}_{\xi D}(\hat{\theta}_n) \approx \widehat{var}_D(\hat{\theta}_n)$$

In addition, if the PSU sample is selected with-replacement or approximately so (when the sampling rate is low as in the CISS), the with-replacement design variance estimator also captures the variance with respect to the model (Binder & Roberts, 2009).

In summary, for the CISS data analysis, design unbiased or nearly design unbiased point estimator can be used to estimate the finite population parameters and the model parameters when the model is correctly specified. The with-replacement design variance estimator captures both design variance and the model variance.

From now on we only consider design unbiased or approximately design unbiased point estimators and their design variance estimators.

4.4 Design-Unbiased Point Estimator

Probability sampling defines a probability space so that the inclusion probability π_k for each sampled unit k can be derived and its inverse $w_k = 1/\pi_k$ can be used to weight the data to obtain (approximately) design unbiased estimators. The design-unbiased point estimator is robust because it is unbiased for the finite population parameter whether the super-population model that generated the finite population is true or not.

Unweighted estimators, on the other hand, may incur severe bias. In Table 1 for example, the unweighted crash distribution by PAR domain estimated from the 2017 CISS sample, which is simply the 2017 CISS sample allocation to the PAR domains predetermined by NHTSA, is severely biased compared with the weighted distribution (Table 2).

4.5 Design Variance Estimation

The impact of the sample design must be recognized when one estimates $Var_D(\hat{\theta}_n)$. Ignoring the sample design may cause severe bias in the standard error estimates.

Estimation methods and computer software have been developed to estimate $Var_D(\hat{\theta}_n)$. Specialized procedures for complex survey data analysis, such as SAS SURVEY procedures and SUDAAN procedures, should be used for CISS data analysis along with proper design statements. Because of the small CISS PSU sampling fractions, the with-replacement design option can be used for CISS data analysis.

Different variance estimation methods (for example, the Jackknife variance estimation method and the Taylor series method) can be used to estimate the standard errors of CISS estimates. See

Wolter (2007) for more information about design variance estimation under a complex sample design.

4.6 Alternative Estimation Methods

Alternative approaches to inference from survey data are not in the scope of this document but can be found in the literature. For more information about complex survey data analysis and other alternative inference approaches, see Chambers and Skinner (2003), Pfeffermann and Rao (2009), Graubard and Korn (1996).

4.7 Missing Data and Imputation

As other surveys, CISS suffers from unit missing and item missing. In the CISS, unit missing refers to sampled crashes that were not investigated because the key vehicle was not available for inspection and thereby was replaced. These cases have little useful information therefore are excluded from the final analysis file. On the other hand, item missing in the CISS refers to individual missing values of study variables of the investigated cases.

The CISS base weights are created to be used with the full original sample, instead of only the investigated cases. Therefore when there are missing PJs or PARs, using only the investigated sample without any treatment to the base weights may result in biased estimates. In the CISS, unit missing are treated by nonresponse adjustment to the base weights (see Chapter 3).

Various methods have been proposed in the literature for item missing treatment – many involving imputation. See Brick and Kalton (1996) for reviews of imputation and weight adjustment methods commonly used. In the CISS data file, item missings are not imputed. Data users should use the method that fits their study to handle the item missings. Two examples are presented in the next chapter.

5. Examples

In this Chapter, we illustrate how to specify design statements in SAS and SUDAAN software under different scenarios and how to handle item missings using SAS PROC SURVEYIMPUTE through the following examples:

- Example 1: Calculating single-year CISS estimates.
- Example 2: Calculating domain estimates with Hot Deck imputation.
- Example 3: Combining multiple years of data.
- Example 4: Fully efficient fractional imputation.

5.1 Example 1: Single-Year CISS Estimates

The following SAS and SAS-callable SUDAAN programs show how design options are specified to make single-year CISS estimates. We choose Jackknife variance estimation method as the variance estimation method in SAS and SAS-callable SUDAAN programs. This also implicitly assumes the PSUs were selected with replacement or with a low sampling rate (as in CISS).

CISS analysis data files include a file (JKWGT) of adjusted Jackknife replicate weights. JK replicate weights were created by deleting one PSU at a time and then recalculate the PSU weights and implement the same weighting process used to create the final case weights. These JK replicate weights capture the effect of weight adjustment and use the finest PSU stratification (12 PSU strata for 2017, 13 PSU strata for 2018, 16 PSU strata for 2019 and afterward) so they may produce better variance estimates. It should be noticed these adjusted JK replicate weights can only be used for single-year estimation. When multiple years of CISS or CDS data are combined for analysis, either Taylor series method or unadjusted JK replicate weights should be used. See more details in Example 3.

The final CISS weight variable, CASEWGT, should be used in a weight statement. JKWGT1 – JKWGT24 are the 24 adjusted JK replicate weights for 2017 CISS. The JKCOEFS=0.5 in SAS statement and ADJJACK=0.5 in SUDAAN statement are associated with 2017 CISS JK replicate weights. For future CISS, these coefficients may be different and will be published in future CISS AUM.

The following examples are the SAS and SUDAAN programs and outputs for single-year estimates of domain identification variable CATEGORY defined in Table 1 using the adjusted JK replicate weights.

The input data file CISS_CRASH is the 2017 CISS CRASH file. It is already merged with the adjusted JK weight file (JKWGT). When JK replicate weights are provided by the user through the REPWEIGHTS statement in SAS or the JACKWGTS statement in SUDAAN, the design statements (the STRATA and the CLUSTER statements in SAS and the NEST statement in SUDAAN) are not needed.

```
/*SAS Example*/
proc surveyfreq data=ciss_crash varmethod=jk;
     format category domainfmt.;
     tables category;
     repweights JKWGT1-JKWGT24 / JKCOEFS=0.5;
     weight CASEWGT;
     run;
```

Table 2: Single-year CISS estimates - SAS Output

CASE CATEGORY					
CATEGORY	**Frequency**	**Weighted Frequency**	**Std Dev of Wgt Freq**	**Percent**	**Std Err of Percent**
Fatal PV Crash	81	15,798	1,361	0.5692	0.0506
Recent-Model PV & Severe	210	24,940	2,130	0.8985	0.0656
Recent-Model PV & Injured	481	312,681	16,639	11.2653	0.5214
Recent-Model PV & No Injury	286	540,462	27,766	19.4718	0.8064
Mid-Model PV & Severe	122	20,745	2,548	0.7474	0.0833
Mid-Model PV & Injured	233	226,985	13,614	8.1778	0.4692
Mid-Model PV & No Injury	219	439,050	21,894	15.8181	0.5946
Old-Model PV & Severe	95	49,779	8,421	1.7934	0.2885
Old-Model PV & Injured	194	453,913	65,750	16.3537	2.4256
Old-Model PV & No Injury	114	691,255	85,055	24.9046	2.5432
Total	2,035	2,775,608	108,648	100.000	

```
/*SAS-Callable SUDAAN Example*/
PROC CROSSTAB DATA=ciss_crash DESIGN=JACKKNIFE NOTSORTED;
     WEIGHT    CASEWGT;
     JACKWGTS  JKWGT1-JKWGT24 / adjjack=0.5;
     TABLES    CATEGORY;
     CLASS     CATEGORY;
     SETENV    LABWIDTH=30 COLWIDTH=15;
     PRINT     NSUM="SAMSIZE" WSUM="POPSIZE" SEWGT="POP SE"
               COLPER="PERCENT" SECOL="PERCENT SE"
               / STYLE=NCHS NSUMFMT=F7.0 WSUMFMT=F8.0
                SEWGTFMT=F8.0;
     RFORMAT   CATEGORY DOMAINFMT.;
     RUN;
```

Table 3: Single-year CISS estimates – SAS-Callable SUDAAN Output

Variance Estimation Method: Replicate Weight Jackknife
by: CASE CATEGORY.

CASE CATEGORY	SAMSIZE	POPSIZE	POP SE	PERCENT	PERCENT SE
Total	2035	2775608	108648	100.00	0.00
Fatal PV Crash	81	15798	1361	0.57	0.05
Recent-Model PV & Severe	210	24940	2130	0.90	0.07
Recent-Model PV & Injured	481	312681	16639	11.27	0.52
Recent-Model PV & No Injury	286	540462	27766	19.47	0.81
Mid-Model PV & Severe	122	20745	2548	0.75	0.08
Mid-Model PV & Injured	233	226985	13614	8.18	0.47
Mid-Model PV & No Injury	219	439050	21894	15.82	0.59
Old-Model PV & Severe	95	49779	8421	1.79	0.29
Old-Model PV & Injured	194	453913	65750	16.35	2.43
Old-Model PV & No Injury	114	691255	85055	24.90	2.54

5.2 Example 2: Calculating Domain Estimates With Hot Deck Imputation

Domain estimate refers to the statistics for a subpopulation. It is important to use the full sample for domain estimation. It may produce biased variance estimate by subsetting the full sample for domain estimation.

In SAS SURVEY procedures, domains are specified by the variables listed in the TABLES and/or the DOMAIN statement. The SAS BY statement subsets the full sample for one domain at a time, therefore it should not be used to produce domain estimates. In SAS-callable SUDAAN procedures, domains are specified by the variables listed in the TABLES and/or the SUBPOPN statement.

The following SAS program estimates the total number and the average number of vehicles involved in 2017 CISS crashes during nighttime (7 p.m. – 6:59 a.m.). We keep all records in the file although we are only interested in nighttime estimates.

Variable NIGHT identifies nighttime (7 p.m. – 6:59 a.m.) and day time (7 a.m. – 6:59 p.m.). Three crashes have missing crash time which caused 3 missing values in variable NIGHT. So we first run SAS PROC SURVEYIMPUTE[3] to impute variable NIGHT. To preserve the observed multivariate relationship between variable NIGHT and the number of vehicles involved, these two variables are imputed jointly using METHOD=HOTDECK along with option SELECTION=WEIGHTED to avoid imputation bias. The domains are defined by variable NIGHT in the DOMAIN statement.

```
PROC SURVEYIMPUTE DATA=CISS_CRASH_A METHOD=HOTDECK
     (SELECTION=WEIGHTED) SEED=13579;
     VAR       NIGHT VEHICLES;
     OUTPUT    OUT=CISS_CRASH_B;
     RUN;


PROC SURVEYMEANS DATA=CISS_CRASH_B VARMETHOD=JK MEAN SUM CLM;
     FORMAT          NIGHT NIGHTFMT.;
     VAR             VEHICLES;
     DOMAIN          NIGHT;
     REPWEIGHTS      JKWGT: / JKCOEFS=0.5;
     WEIGHT          CASEWGT;
     RUN;
```

[3] For more details see: https://support.sas.com/documentation/onlinedoc/stat/141/surveyimpute.pdf

Table 4: SAS PROC SURVEYMEANS domain estimates

Statistics for Night Domains						
Night	Mean	Std Error of Mean	95% CL for Mean		Sum	Std Error of Sum
Night	1.501198	0.043321	1.411789	1.590608	1343701	92199
Day or Unknown	1.881356	0.028691	1.82214	1.940572	3537932	167199

5.3 Example 3: Combining Multiple Years of Data

Combining multiple years of data increases sample size for better estimates. In this section, we explain the issue and show how to use multiple years of CISS/CDS data. In the following, we first explain how to combine multiple years of CISS data and then explain how to combine multiple years of CDS data with multiple years of CISS data.

Because it takes several months to train a CISS data collection technician, CISS data collection sites (PSUs) were phased in over two years: 24 PSUs in 12 strata collecting data from January 2017, 28 PSUs in 13 strata collecting data from January 2018, and 32 PSUs in 16 strata collecting data from July 2018. Here the 24 PSU sample is a sub-sample of the 28 PSU sample and the 28 PSU sample is a sub-sample of the 32 PSU sample. Therefore, these annual PSU samples are not independent samples. In addition, the strata of the 24 PSU sample were collapsed from the strata of the 28 PSU sample and the strata of the 28 PSU sample were collapsed from the strata of the 32 PSU sample.

The adjusted JK replicate weights were created for single-year estimates, not for combined multiple year data analysis. For multiple year data analysis, Taylor series method or the unadjusted JK replicate weights should be used. The unadjusted JK replicate weights refer to the JK replicate weights that are not adjusted by the weighting procedures – directly generated by SAS procedures for example. Taylor series method and the Jackknife method using the unadjusted JK replicate weights do not capture the weighting adjustment effects so they may produce larger variance estimates.

When multiple years of CISS data are combined, the coarsest PSU strata are used for variance estimation. The 2017 CISS has the coarsest PSU strata (12 strata) compared with 13 and 16 strata in 2018 and later years. Therefore, the 12 PSU strata of 2017 CISS are used for variance estimation whenever multiple years of data are combined. In each year's CISS data files, a variable PSUSTRAT identifies the 12 PSU strata for variance estimation purpose and variable PSU identifies the PSU. In the following SAS program, we first combine 2017 - 2018 CISS data into a file "COMBINED_CISS" in a SAS data step. The first SAS PROC SURVEYFREQ uses Taylor series method to estimate the variances. The second SAS PROC SURVEYFREQ generates the unadjusted JK replicate weights and uses these unadjusted JK replicate weights to estimate the variance.

```
DATA COMBINED_CISS;
     SET CISS_2017 CISS_2018;
     RUN;

/*Taylor series method*/
PROC SURVEYFREQ DATA=COMBINED_CISS VARMETHOD=TAYLOR;
     STRATA PSUSTRAT;
     CLUSTER PSU;
     FORMAT CATEGORY DOMAINFMT.;
     TABLES CATEGORY;
     WEIGHT CASEWGT;
     RUN;

/*Using unadjusted JK replicate weights generated by SAS*/
PROC SURVEYFREQ DATA=COMBINED_CISS VARMETHOD=JK;
     STRATA PSUSTRAT;
     CLUSTER PSU;
     FORMAT CATEGORY DOMAINFMT.;
     TABLES CATEGORY;
     WEIGHT CASEWGT;
     RUN;
```

CISS target population (crashes with at least one passenger vehicle towed) is larger than CDS target population (crashes with at least one passenger vehicle towed due to damage or unknown reason). When comparable sub-populations can be identified in both CISS and CDS, multiple years of data from both surveys can also be combined. In 2019, NHTSA added a vehicle level variable TOWED. In 2019 CISS data file, variable TOWED=2 if the vehicle was towed due to disabling damage or towed due to unknown reason. Therefore, the following SAS statement can be used to identify passenger vehicle towed due to damage or unknown reason from 2019 CISS file:

```
IF TOWED IN (2) AND 1<=BODYTYPE<=49 THEN PV_TOW_DAMAGE=1;
```

Here variable BODYTYPE is the vehicle body type.

NHTSA plans to add a new category 7 - “Towed, Unknown Reason” to variable TOWED from 2020. Therefore, the following SAS statement can be used to identify passenger vehicle towed due to damage or unknown reason from 2020 CISS file:

```
IF TOWED IN (2,7) AND 1<=BODYTYPE<=49 THEN PV_TOW_DAMAGE=1;
```

Variable PV_TOW_DAMAGE is used to create a crash level flag to identify CDS in-scope crashes in CISS: e.g. CDS_IN_SCOPE=1 if there is at least one vehicle in the crash has PV_TOW_DAMAGE=1 and CDS_IN_SCOPE=2 otherwise. Variable CDS_IN_SCOPE is set to 1 for all crashes in CDS. Then the full CISS sample (CDS in-scope or not) can be combined with

the CDS full sample and variable CDS_IN_SCOPE variable is used as a domain identifier for combined data analysis.

CISS sample selection is independent from CDS sample selection. When CISS data is combined with CDS data, a new stratification variable STUDY (STUDY=1 for CDS and STUDY=2 for CISS) is created first in a SAS data step to capture this independence. In the following example, 2 years of CDS data are combined with 2 years of CISS data:

```
DATA CDS_CISS;
     SET  CDS2014.ACCIDENT (IN=CDS2014)
          CDS2015.ACCIDENT (IN=CDS2015)
          CISS2019.CRASH  (IN=CISS2019)
          CISS2020.CRASH  (IN=CISS2020);
     STUDY = CDS2014*1 + CDS2015*1 + CISS2019*2 + CISS2020*2;
     … … … …
     RUN;
```

Since multiple years of data are combined, we use Taylor series method or the unadjusted JK replicate weights to estimate the variances. Variable STUDY is used as an extra stratification variable. The PSU identification variable, PSU, now is the third variable listed in the SUDAAN NEST statement (PSULEV=3). If we let the software generate the JK replicate weights, the design statements (the STRATA and the CLUSTER statements in SAS and the NEST statement in SUDAAN) must be included. Variable CDS_IN_SCOPE is used as a common domain identifier.

```
/*SUDAAN Example*/
PROC CROSSTAB DATA=CDS_CISS FILETYPE=SAS DESIGN=JACKKNIFE
              NOTSORTED;
     NEST      STUDY PSUSTRAT PSU / PSULEV=3;
     SUBPOPN   CDS_IN_SCOPE=1 / NAME="CDS IN-SCOPE ONLY";
     WEIGHT    CASEWGT;
     … … …
     RUN;

/*SAS Example*/
proc surveyfreq DATA=CDS_CISS VARMETHOD=TAYLOR;
     STRATA    STUDY PSUSTRAT;
     CLUSTER   PSU;
     WEIGHT    CASEWGT;
     DOMAIN    CDS_IN_SCOPE;
     … … …
     run;
```

5.4 Example 4: Fully Efficient Fractional Imputation

When item missing rate is high, Hot Deck imputation may incur non-negligible imputation variance. The fully efficient fractional imputation method (FEFI) preserves the observed multivariate relationship and does not add additional variability due to imputing value selection (Kim and Fuller, 2004). FEFI method distributes the weights of the units with missing items to multiple units with observed items by adjusting the final weights and the replicate weights to compensate the missing items. So it must be used with replicate variance estimation methods.

In the following example, we illustrate how to impute multiple variables jointly using SAS PROC SURVEYIMPUTE and use the output file for analysis.

First, in the following SAS data step, three variables with missing values are coded:

- INJURED (indicator for injury or fatal crash): 1 = yes, 0 = no, 9 is set to SAS missing.
- NIGHT (indicator for night crash): 1 = night, 0 = day, 99 is set to SAS missing.
- ALCINV (indicator for alcohol involved crash): 1 = yes, 0 = no, 9 is set to SAS missing.

```
DATA CISS_CRASH_C;
     SET CISS_CRASH;
     IF SUBSTR(CRASHTIME,1,2) IN ('19','20','21','22','23','00',
     '01','02','03','04','05','06') THEN NIGHT=1;
     ELSE IF SUBSTR(CRASHTIME,1,2)='99' THEN NIGHT=.;
     ELSE NIGHT=0;
     IF CAIS=9  THEN INJURED=.; ELSE INJURED=(1<=CAIS<=7);
     IF ALCINV=9 THEN ALCINV =.; ELSE ALCINV=(ALCINV=1);
     RUN;
```

Then SAS PROC SURVEYIMPUTE is used to impute the missing values of the three variable simultaneously to preserve the observed multivariate relationship using FEFI method. As the result, 383 records have at least one of the three variables missing and all missing items are imputed.

```
PROC FORMAT;
     VALUE YESNOUNK17F
     1 = 'Yes'
     2 = 'No';
     RUN;

PROC SURVEYIMPUTE DATA=CISS_CRASH_C METHOD=FEFI;
     FORMAT INJURED NIGHT ALCINV YESNOUNK17F.;
     REPWEIGHTS JKWGT: ;
     WEIGHT   CASEWGT;
```

```
VAR      INJURED ALCINV NIGHT;
CLASS     INJURED ALCINV NIGHT;
ID      CASEID;
OUTPUT    OUT=INJURED_FEFI;
RUN;
```

Table 5: FEFI imputation output

Imputation Summary		
Observation Status	**Number of Observations**	**Sum of Weights**
Nonmissing	1,652	2237242.4
Missing	383	538365.588
Missing, Imputed	383	538365.588
Missing, Not Imputed	0	0
Missing, Partially Imputed	0	0

FEFI method modifies the final weights and the user provided replicate weights. The new imputation-adjusted weights are output to data file INJURED_FEFI. The new final weight is named as IMPWT. The new replicate weights are named as IMPREPWT_#. These imputation-adjusted weights must be used in the analysis. The following is a SAS example:

```
PROC SURVEYLOGISTIC DATA=INJURED_FEFI VARMETHOD=JK;
     CLASS     ALCINV NIGHT;
     MODEL     INJURED = ALCINV NIGHT;
     WEIGHT    IMPWT;
     REPWEIGHTS IMPREPWT_: / JKCOEFS=0.5;
     RUN;
```

6. Frequently Asked Questions

1. What is CISS?

A. The Crash Investigation Sampling System is NHTSA's new national probability-based crash sampling system designed to replace the Crashworthiness Data System.

2. What data does CISS collect and what does it represent?

A. CISS selects crash sample by selecting police crash reports and collects information about the crash, event, passenger and vehicle. The CISS data, when used with the accompanying weights, are nationally representative of all police-reported motor vehicle traffic crashes on a traffic way where at least one passenger vehicle is towed.

3. When did NHTSA transition from CDS to CISS?

A. 2015 was the last year of data collection through CDS. CISS was designed and implemented over a multi-year effort and started collecting data for publication in January 2017.

4. Why did NHTSA transition from CDS to CISS?

A. The CDS had used the same data collection sites since 1988. Over the years, the population has shifted, the vehicle fleet and the analytic needs of the safety community have changed. The existing CDS police jurisdiction samples and weights became outdated as the PJ population changed. Congress directed and provided funds to NHTSA to modernize its data collection system.

5. How is CISS different from CDS in terms of its sample design?

A. The following are some major differences in sample designs of CISS and CDS:

- Independent samples: The CISS sample design is independent from any other NHTSA's surveys, including NHTSA's new Crash Report Sampling System that replaces the NASS General Estimates System. In comparison, the GES and the CDS samples were nested, i.e., the CDS used a subset of the GES data collection sites. The independent design allows NHTSA to optimize each system - CISS and CRSS.
- Different formation of PSUs: In both CISS and CDS, a PSU is either a county or a group of counties. In CISS, the nation was partitioned into 1,784 PSUs, while in CDS 1,195 PSUs were formed. CISS's average PSU size is smaller than CDS. This resulted in more operationally efficient PSUs in CISS. In addition, a new composite PSU measure of size variable using the various estimated crash counts by the new PAR domains was used in CISS.
- Scalable PSU sample: The CISS PSU sample size can be increased without changes to the existing PSU sample while the corresponding selection probabilities are still trackable. This enables NHTSA to accommodate potential budget fluctuations with minimum operational costs and efforts.
- Scalable PJ sample: The Pareto sampling method was used to select the CISS PJ sample. The second stage sampling frame, the police jurisdictions in the selected PSUs, changes over time. Consequently, the PJ sample needs to be reselected

occasionally to maintain adequate sample size or to cover the updated PJ frame. Pareto sampling reduces the changes to the existing PJ sample when a new PJ sample is reselected.

- Alignment with data needs: PAR domains were revised based on data needs to oversample crashes involving killed or injured occupant. At the PAR sample selection stage, PAR domains are used to oversample high interest crashes.
- Optimized sample allocation: CISS PSU, PJ, and PAR sample sizes were determined by minimizing the variance of a simplified variance estimator subject to fixed cost.
- Replacement cases: In the CISS, if the vehicle that defines the PAR domain is not available for investigation, replacement case is selected and investigated. This new feature results in about 16 percent more useful cases in 2017.
- Weight adjustments: In the CISS, non-responding PJs and PARs are monitored and weight adjustments are applied to mitigate potential bias. In addition, large weights are truncated by the 10 PAR domains.
- Jackknife replicate weights for variance estimation: Adjusted Jackknife replicate weights are provided in one of the CISS analysis files for variance estimation. These adjusted Jackknife replicate weights capture the impact of the weight adjustments to the total variance.

6. How is CISS similar to CDS in sample design?

A. The following are some major common features between the sample designs of CISS and CDS:

- Both CISS and CDS have a three-stage sample design: PSU, PJ, and PAR sample selection.
- In both surveys, PSUs, PJs and PARs have selection probabilities proportional to their measure of sizes.
- In both surveys, PAR samples are selected weekly by PSU.
- Both surveys collect accident, even, vehicle, and occupant level information.

7. How do the CISS analysis files (data sets) differ from the CDS?

A. CISS analytic datasets differ significantly from the CDS datasets. In CDS, variables were stored in 11 files by the entities (crash, vehicle, person, etc.) from which the variables were collected, while in CISS variables were stored in 37 files by the data modules/tables (airbag, injury, fire, etc.) from which the variables were collected. In addition, the adjusted Jackknife replicate weights were also provided in a separate file. Through the modernized infrastructure, CISS has collected and is presenting much more information from its crash investigations (including EDR information, child safety information and scene diagram coordinates). Variable names have also changed between CDS and CISS. CISS data users should refer extensively to the CISS and CDS AUMs.

8. How should data users compute the variance for CISS estimates?

A. The CISS sample is the result of complex survey sampling, and therefore is not a simple random sample. Software specialized in complex survey data analysis such as SAS SURVEY procedures or SUDAAN procedures should be used to make estimates from CISS sample. Using these specialized softwares along with the appropriate design and weight statements, the sampling variance can be estimated. Failing to take the sample

design and weights into account in estimation may incur severe bias to the point and variance estimates. NHTSA created Jackknife replicate weights for the CISS variance estimation using single-year data. These Jackknife replicate weights incorporate weighting adjustments therefore they capture the effects of these weighting adjustment. See Chapters 4 and 5 for some basic concept of complex survey data analysis, and SAS and SUDAAN examples on how to estimate the variances for CISS estimates.

9. What software or techniques should be used for variance calculation?

A. Any software that takes complex survey design into account can be used to make estimates from CISS sample. Some examples of such softwares: SAS SURVEY procedures, SUDAAN, R survey package, and STATA. See Chapter 5 for specific SAS and SUDAAN examples of programming techniques for variance estimation.

10. Are producing small area estimates different from CDS to CISS?

B. The problem associated with a small sample size does not change from CDS to CISS. Users can combine multiple years of CISS data together to augment the sample size. See Chapter 5 for more details on how to specify design options when multiple years of CISS data are combined. Also see Rao and Molina (2015) for more details on small area estimation.

11. How are missing data addressed in CISS?

A. NHTSA made nonresponse adjustments to the weights to treat the unit missings in CISS. To handle item missings in CISS, data users need to choose their own methods for their study. See section 4.7 and Chapter 5 for more details and some examples.

12. Can CISS data be combined with CDS data?

B. Multiple years of CISS data and CDS data can be combined if comparable sub-populations can be identified in both CISS and CDS because CISS target population is different from CDS target population. See section 5.3 for more detailed discussion and example.

7. References

Binder, D. A., & Roberts, G. (2009). Design- and model-based inference for model parameters. In D. Pfeffermann & C. R. Rao, *Handbook of Statistics* 29 Vol. 29B Amsterdam: North-Holland [Elsevier].

Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research, Vol. 5*, pp 215-238.

Chambers, R. L. & Skinner, C. J. (2003). *Analysis of survey data.* John Wiley & Sons Ltd.

Graubard, B. I. & Korn, E. L. (1996). Survey inference for subpopulations. American Journal of Epidemiology, Vol. 144, No. 1, pp 102-106.

Graubard, B. I. & Korn, E. L. (1996). Modeling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research, Vol. 5*, No. 3, pp 263-282.

Hartley, H. O. & Sielken, R. L. (1975). A "super-population viewpoint" for finite population sampling. *Biometrics, Vol. 31*, No. 2, pp 411-422.

Kim, J. K. & Fuller, W. A. (2004). Fractional Hot Deck Imputation. *Biometrika, Vol. 91,* pp 559-578.

Pfeffermann, D., & Rao, C. R. (2009). Handbook of statistics 29, Vol. 29B. Amsterdam: North-Holland [Elsevier].

Rao, J. N. K., & Molina, I. (2015). *Small area estimation*. New York: Wiley Series in Survey Methodology.

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference, Vol. 62*, pp. 159-191.

SAS/STAT 14.1 User's Guide: The SURVEYIMPUTE Procedure. Cary, NC: SAS Institute.

Wolter, K. (2007). *Introduction to variance estimation*. New York: Springer-Verlag New York, Inc.

Zhang, F., Noh, E. Y., Subramanian, R., & Chen, C-L. (In press). Crash Investigation Sampling System: Sample design and weighting. Washington, DC: National Highway Traffic Safety Administration.

DOT HS 812 801
September 2019

U.S. Department
of Transportation

**National Highway
Traffic Safety
Administration**

14418-082219-v3