U.S. Department
of Transportation

National Highway
Traffic Safety
Administration

NHTSA

DOT HS 812 509

March 2018

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

DISCLAIMER

This publication is distributed by the U.S. Department of Transportation, National Highway Traffic Safety Administration, in the interest of information exchange. The opinions, findings, and conclusions expressed in this publication are those of the authors and not necessarily those of the Department of Transportation or the National Highway Traffic Safety Administration. The United States Government assumes no liability for its contents or use thereof. If trade or manufacturers' names or products are mentioned, it is because they are considered essential to the object of the publication and should not be construed as an endorsement. The United States Government does not endorse products or manufacturers.

Suggested APA Format Citation:

Zhang, F., Subramanian, R., Chen, C.-L., & Noh, E. Y. (2018, March). *Crash Report Sampling System: Design overview, analytic guidance, and FAQs* (Report No. DOT HS 812 509). Washington, DC: National Highway Traffic Safety Administration.

Technical Report Documentation Page

1. Report No. DOT HS 812 509	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Crash Report Sampling System: Design Overview, Analytic Guidance, and FAQs		5. Report Date March 2018
		6. Performing Organization Code NSA-210
7. Authors 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 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

Abstract

This document describes the Crash Report Sampling System 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 CRSS data analysis.

17. Key Words NHTSA, CRSS, GES, NASS, sample design, complex survey data analysis, analytic guidance.	18. Distribution Statement This document is available through 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 35	22. Price

Acronyms

- 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
- 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

Acronyms ii

1. **Introduction** 1

2. **The CRSS Sample Design** 2

3. **CRSS Weighting Procedures** 5

4. **Basic Concepts of Complex Survey Data Analysis** 6

4.1 Model Parameter Estimation *6*

4.2 Finite Population Parameter Estimation *7*

4.3 Two-Step Sampling Procedure *8*

4.4 Design-Unbiased Point Estimator *10*

4.5 Design Variance Estimation *10*

5. **Estimation Examples** 11

5.1 Example 1: Single Year CRSS Estimates Using SAS and SUDAAN *11*

5.2 Example 2: Combining Multiple Years of GES and CRSS Data *14*

5.3. Example 3: Composite Estimates *20*

5.4 Example 4: Domain Estimates *21*

6. **Frequently Asked Questions** 23

7. **References** 29

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 data collection. Given the shifts in population and the vehicle fleet, and the changing analytic needs of the safety community, the U.S. Congress authorized NHTSA to modernize its crash data collection system.

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

This document provides an overview of the CRSS sample design and weighting procedure, and describes some basic concepts on analysis of complex survey data. In addition, it provides examples to show how to make estimates using CRSS and GES data and discusses issues related to CRSS data analysis. Finally, this document catalogs frequently asked questions on sampling and estimation as they apply to GES/CRSS and the Fatality Analysis Reporting System.

While this document provides a broad overview of the design of CRSS, a supplemental NHTSA technical report, *Crash Record Sampling System: Sample Design and Estimation* (Zhang, Noh, Subramanian, & Chen, in press) to be published by NHTSA will illustrate the CRSS sample design and weighting procedures in greater detail.

2. The CRSS Sample Design

CRSS was designed independent of other NHTSA surveys. The target population for the CRSS is the same as that for the GES: all police-reported motor vehicle crashes on trafficways. Because a nationwide direct selection of PARs is currently infeasible, the CRSS PAR sample is selected in multiple stages to produce a nationally representative probability sample.

At the first stage of selection, 3,117 counties in the United States were grouped into 707 primary sampling units. A PSU in the CRSS is either a county or a group of counties. U.S. territories, some remote counties in Alaska, and small islands of Hawaii were excluded.

The 707 PSUs in the PSU frame (the collection of all PSUs) were stratified into 50 strata by the four Census regions, urban/rural, vehicle miles traveled, total number of crashes, total truck miles traveled, and road miles. Each of the 707 PSUs in the frame was assigned a measure of size (MOS) equal to the combination of its estimated nine types of crash counts defined in Table 1 below. First, 101 PSUs were selected by a stratified probability proportional-to-size sampling method. Then a sequence of sub-samples was selected from the 101 PSU sample, and in this process the strata were collapsed if necessary. This produced a sequence of nested PSU samples with decreasing sample sizes selected from the collapsed strata. These nested PSU samples allow NHTSA to change the PSU sample size without reselecting the sample in the future. Therefore, the final PSU sample was the result of a multiphase sampling, and the PSU sample was selected in such a way that the resulting selection probability was still approximately PPS.

For the 2016 CRSS, 60 PSUs were selected from 24 PSU strata. Due to the non-response of 7 PSUs, CRSS data were collected from 53 PSUs. A PSU level non-response adjustment was applied to mitigate the potential non-response bias.

The secondary sampling units are police jurisdictions. Within each selected PSU, PJs were stratified into three PJ strata by their estimated measure of size which is a combination of crash counts in six categories of interest. The Pareto sampling method 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 if a new PJ sample would need to be selected because of PJ frame (the collecton of all PJs in the selected PSU) changes. The PJ inclusion probability under the Pareto sampling is approximately PPS (Rosén, 1997). Across the 53 responding PSUs, a total of 350 PJs was selected and 337 PJs cooperated in 2016. Weight adjustments were made to mitigate the potential bias caused by the 13 non-responding PJs.

The tertiary sampling units are PARs. CRSS Data collectors periodically receive PARs from the selected PJs. All new PARs are sequentially stratified in the order they become available into nine PAR strata (see Table 1 below). These nine PAR strata were formed based on the results of NHTSA's internal data needs and public data needs studies. The PAR stratification is used to over-sample the following important analysis domains to ensure enough cases are selected into the sample:

- Crashes involving killed or injured pedestrians;
- Crashes involving killed or injured motorcycle occupants;
- Crashes involving killed or injured occupants in a late model year passenger vehicle; and
- Crashes involving killed or severely injured occupants in a non-late-model-year passenger vehicle.

From each PAR stratum, a systematic sampling method is used to select the PAR sample. The sampling intervals are determined in such a way that the final weights are approximately equal for all the PARs in the same PAR stratum in order to reduce the sampling variance for the domain estimates. The target PAR sample size is around 50,000 every year.

See *Crash Record Sampling System: Sample Design and Estimation*, (Zhang, Noh, Subramanian, & Chen, in press) for more detailed information on CRSS sample design.

Table 1: 2016 CRSS Weighted and Unweighted Estimates of Crash Distribution by PAR Strata

CRSS PAR Strata	PAR Strata Description	Unweighted Distribution[1] (percent) Resulting Sample Allocation	Unweighted Standard Error (percent)	Weighted Distribution (percent)	Weighted Standard Error (percent)
2	Crashes not in Stratum 1 that: • Involves a killed or injured (includes injury severity unknown) non-motorist	8.26	0.13	2.11	0.19
3	Crashes not in Stratum 1 or 2 that: • Involves a killed or injured (includes injury severity unknown) motorcycle or moped rider	5.00	0.10	1.31	0.08
4	Crashes not in Stratum 1-3 that: • At least one occupant of a late model year[2] passenger vehicle is killed or incapacitated	2.87	0.08	0.42	0.04
5	Crashes not in Stratum 1-4 that: • At least one occupant of an older[3] passenger vehicle is killed or incapacitated	6.33	0.11	1.50	0.15
6	Crashes not in Stratum 1-5 that: • at least one occupant of a late model year passenger vehicle is injured (including injury severity unknown)	15.70	0.17	7.39	0.39
7	Crashes not in Stratum 1-6 that: • involved at least one medium or heavy truck or bus (includes school bus, transit bus, and motor coach) with GVWR 10,000 lbs. or more	5.48	0.10	6.26	0.18
8	Crashes not in Stratum 1-7 that: • at least one occupant of an older passenger vehicle is injured (including injury severity unknown)	13.36	0.16	14.89	0.43
9	Crashes not in Stratum 1-8 that: • involved at least one late model year passenger vehicle, AND • No person in the crash is killed or injured	22.51	0.19	29.10	0.86
10	Crashes not in Stratum 1-9: This includes mostly PDO* crashes involving a non-motorist, motorcycle, moped, and passenger vehicles that are not late model year and any crashes not classified in strata 1-9.	20.50	0.19	37.03	0.82

*: PDO: Property damage only.

[1] The unweighted estimates ignored the sample design and the weights.

[2] Late model year vehicle: vehicle that is no more than 4 year old.

[3] Older vehicle: vehicle that is more than 4 year old.

3. CRSS Weighting Procedures

The CRSS sample is the result of probability sampling featuring stratification, clustering, and selection with unequal probabilities. Because of these features, the CRSS sample is not a simple random sample and users need to use proper weights to produce unbiased and robust estimates. The 2016 CRSS weights were created as follows:

- Calculate the base weights (the inverse of selection probabilities) at all three stages (PSU, PJ, and PAR).
- Adjust the base weights for non-response[4] at all three stages to correct potential non-response bias.
- Calibrate the PJ and the PAR weights using the PSU level total PAR stratum counts to further correct potential non-response bias and coverage bias.
- Adjust the weights for duplicates.

The final weight variable for the CRSS estimation is WEIGHT. See , (Zhang, Noh, Subramanian, & Chen, in press) for more detailed information on the CRSS weighting procedure.

[4] Non-responding PARs are incomplete PARs or non-readable PARs. Non-responding PJs and PSUs are PJs and PSUs refused to cooperate.

4. Basic Concepts of Complex Survey Data Analysis

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 $\{y_1, y_2, \dots, y_N\}$:

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

observed from the N crashes reported in the year 2016. One may view these observations as outcomes of independent and identical Bernoulli trials indexed by model parameter θ:

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

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:

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

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 at year 2016. Similar statistics include N (2016 total number of crashes) and $\sum_{k=1}^{N} y_k$ (2016 total number of fatal 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 2016 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 the 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 PARs was selected from a finite population of PARs for CRSS data collection and how the final CRSS 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 was too costly to obtain all PARs in the US to directly select a PAR sample. PARs 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 assign different selection probabilities. 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 2016 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 and 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:

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

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 summary, a design unbiased or nearly design unbiased point estimator can be used to estimate the finite population parameter regardless if the super-population model is correct or not. If the super-population model is correctly specified and the finite population parameter is unbiased with respect to the model for the model parameter, then the design unbiased estimator can also be used to estimate the model parameter. The design unbiased variance estimator for the design unbiased point estimator not only can be used to estimate the design variance of the design unbiased estimator, but also can be used to estimate its total variance when the finite population size is large.

In most of our analysis, the finite population size N is indeed very large, therefore, 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 strata estimated from the 2016 CRSS sample, which is simply the 2016 CRSS sample allocation to the PAR strata, is quite different from the weighted distribution, which is an unbiased estimate of the actual crash distribution of all 2016 crashes by PAR strata.

4.5 Design Variance Estimation

The impact of the sample design must be recognized when one estimates $Var_D\left(\hat{\theta}_n\right)$. In Table 1, the unweighted standard errors ignored weights and the sample design. The weighted standard errors take the sample design (including the weights) into account. Table 1 shows ignoring the sample design may cause severe bias to the standard error estimates too.

Estimation methods and computer software have been developed to estimate $Var_D\left(\hat{\theta}_n\right)$. Specialized procedures for complex survey data analysis, such as SAS PROC SURVEY procedures and SUDAAN procedures, should be used for CRSS data analysis along with proper design statements. Because of the small CRSS PSU sampling fractions, the with-replacement design option can be used for CRSS 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 CRSS estimates. We choose to use Jackknife variance estimation method because our simulation study indicates it produces less biased variance estimates for small domain estimates. See Wolter (2007) for more information about design variance estimation under a complex sample design.

5. Estimation Examples

The following examples demonstrate how to use SAS or SUDAAN to calculate CRSS estimates.

- Example 1: Single-year CRSS estimates using SAS and SUDAAN.
- Example 2: Combining multiple years of GES and CRSS data: year-to-year comparisons and significance tests using SUDAAN.
- Example 3: Composite estimates by combining estimates from FARS and CRSS.
- Example 4: Domain estimates using SAS and SUDAAN.

5.1 Example 1: Single Year CRSS Estimates Using SAS and SUDAAN

The following SAS and SAS-callable SUDAAN programs show how design options are specified to make single year CRSS 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 (in our case) with a low sampling rate. We let the software to generate the Jackknife replicate weights.

Variable *PSUSTRAT* defines the PSU strata, and *PSU_VAR* identifies PSUs for variance estimation purpose. In the 2016 CRSS, seven PSUs did not cooperate (refused NHTSA's access to their crash reports). This left some PSU strata with only one responding PSU. In the variable *PSUSTRAT*, these single PSU strata were collapsed with other strata to ensure at least two PSUs per stratum for variance estimation. Also, at the CRSS PSU sampling stage, one PSU was selected with certainty because of its large number of crashes. A certainty PSU is in fact a stratum therefore it is treated as a stratum in *PSUSTRAT*. Variable *PSU_VAR* identifies sampled PSUs. The PJs selected in the certainty PSU are treated as PSUs in *PSU_VAR*.

The final CRSS weight variable, *WEIGHT*, should be used in a weight statement. The input data file IMPUTED.ACCIDENT is the 2016 CRSS crash record data file with imputed variables. The following are the SAS and SUDAAN programs and major outputs.

```
/*SAS Example*/
PROC SURVEYFREQ DATA=IMPUTED.ACCIDENT VARMETHOD=JK;
     STRATA PSUSTRAT;
     CLUSTER PSU_VAR;
     TABLES MAXSEV_IM;
     WEIGHT WEIGHT;
     FORMAT MAXSEV_IM MAXSEV.;
     RUN;
```

Table 2: Single year CRSS estimates - SAS Output:

IMPUTED MAXIMUM INJURY IN CRASH					
MAXSEV_IM	**Frequency**	**Weighted Frequency**	**Std Dev of Wgt Freq**	**Percent**	**Std Err of Percent**
No Injury	22,173	5,061,234	322,636	69.5558	1.0245
Possible Injury	11,225	1,225,708	92,157	16.8447	0.8465
Minor Injury	7,837	736,922	59,854	10.1274	0.5857
Serious Injury	4,971	182,389	18,669	2.5066	0.1992
Fatal	965	34,415	3,141	0.4730	0.0389
Injured, Unknown	326	32,181	13,723	0.4423	0.1915
Died before Crash	2	128.43840	95.82211	0.0018	0.0013
No Person Involved	16	3,526	942.05003	0.0485	0.0124
Total	47,515	7,276,505	438,260	100.000	

```
/*SAS-Callable SUDAAN Example*/
PROC CROSSTAB DATA=IMPUTED.ACCIDENT DESIGN=JACKKNIFE NOTSORTED;
     NEST      PSUSTRAT PSU_VAR;
     WEIGHT    WEIGHT;
     TABLES    MAXSEV_IM;
     CLASS     MAXSEV_IM;
     SETENV    ROWWIDTH=12 COLWIDTH=12 LABWIDTH=12;
     PRINT     NSUM="SAMSIZE" WSUM="POPSIZE" SEWGT="POP SE"
               / NSUMFMT=F6.0 WSUMFMT=F8.0 SEWGTFMT=F8.0;
     RFORMAT   MAXSEV_IM MAXSEV.;
     RUN;
```

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

```
                                      IMPUTED MAXIMUM INJURY IN CRASH.

------------------------------------------------------------------------------------------
|            |                 | IMPUTED MAXIMUM INJURY IN CRASH                          |
|            |                 |----------------------------------------------------------|
|            |                 | Total     | No        | Possible | Minor     | Serious   |
|            |                 |           | Injury    | Injury   | Injury    | Injury    |
------------------------------------------------------------------------------------------
|            |                 |           |           |          |           |           |
|            | SAMSIZE         |    47,515 |    22,173 |    11,225 |     7,837 |     4,971 |
|            | POPSIZE         |7,276,505 |5,061,234 |1,225,708 |   736,922 |   182,389 |
|            | POP SE          |   438,260 |   322,636 |    92,157 |    59,854 |    18,669 |
------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------
|            |                 | IMPUTED MAXIMUM INJURY IN CRASH               |
|            |                 |-----------------------------------------------|
|            |                 | Fatal     | Injured,  | Died      | No        |
|            |                 |           | Unknown   | before    | Person    |
|            |                 |           |           | Crash     | Involved  |
-------------------------------------------------------------------------------
|            |                 |           |           |           |           |
|            | SAMSIZE         |       965 |       326 |         2 |        16 |
|            | POPSIZE         |    34,415 |    32,181 |       128 |     3,526 |
|            | POP SE          |     3,141 |    13,723 |        96 |       942 |
-------------------------------------------------------------------------------
```

5.2 Example 2: Combining Multiple Years of GES and CRSS Data

Combining multiple years of data allows us to make year-to-year comparisons and make better small domain estimates. In this example, multiple years of GES data are combined with 2016 CRSS data. The same approach is also applicable to combining multiple years of GES data with multiple years of CRSS data.

Over the years, NHTSA has noticed some GES estimates might be biased. For example, the 2015 GES fatal crash estimate (21,255 with standard error 1,474) is significantly lower than the 2015 FARS fatal count (32,166). It should be noted the difference between a CRSS estimate and a biased GES estimate may be confounded by the GES bias. To see this, consider the following bias model for the 2015 GES total crash estimate:

$$\hat{t}_{2015}^{GES} = t_{2015} * (1 + B_{2015}^{GES}) * e_{2015}^{GES}$$

where:

- $\hat{t}_{2015}^{GES}$ is the 2015 GES total crash estimate;
- t_{2015} is the 2015 true total crash count;
- e_{2015}^{GES} is the multiplicative error term: $E(e_{2015}^{GES}) = 1$;
- $B_{2015}^{GES} = [E(\hat{t}_{2015}^{GES}) - t_{2015}]/t_{2015}$ is the relative bias of the 2015 GES total crash estimate.

Also, assume the 2016 CRSS total crash estimate is unbiased and:

$$\hat{t}_{2016}^{CRSS} = t_{2016} * e_{2016}^{CRSS}$$

where:

- $\hat{t}_{2016}^{CRSS}$ is the unbiased 2016 CRSS total crash estimate;
- t_{2016} is the 2016 true total crash count;
- e_{2016}^{CRSS} is the multiplicative error term: $E(e_{2016}^{CRSS}) = 1$.

The percent change $\hat{p}$ in the total crash estimate from 2015 to 2016 can be expressed as:

$$\hat{p} = \frac{\hat{t}_{2016}^{CRSS}}{\hat{t}_{2015}^{GES}} - 1$$

Therefore,

$$E(\hat{p}) \approx \frac{t_{2016}}{t_{2015} * (1 + B_{2015}^{GES})} - 1 \neq \frac{t_{2016}}{t_{2015}} - 1$$

It can be inferred that the percent change is confounded by the relative bias.

Similarly, the change $\hat{d}$ in total crash estimates from 2015 to 2016 is:

$$\begin{aligned}\hat{d} &= \hat{t}_{2016}^{CRSS} - \hat{t}_{2015}^{GES} \\ &= t_{2016} * e_{2016}^{CRSS} - t_{2015} * (1 + B_{2015}^{GES}) * e_{2015}^{GES}\end{aligned}$$

Therefore,

$$E(\hat{d}) = (t_{2016} - t_{2015}) - t_{2015}B_{2015}^{GES} \neq t_{2016} - t_{2015}$$

In other words, if GES estimate is biased, then the change from GES estimate to CRSS estimate is confounded by the GES bias. A significance test on the difference estimate $\hat{d}$ is testing whether $(t_{2016} - t_{2015}) - t_{2015}B_{2015}^{GES}$ is zero instead of testing whether $t_{2016} - t_{2015}$ is zero. This demonstrates why comparisons between CRSS and GES should be performed with caution.

However, the comparison among GES estimates is less likely to be confounded by the potential GES bias. To see this, let:

$$\hat{t}_{2014}^{GES} = t_{2014} * (1 + B_{2014}^{GES}) * e_{2014}^{GES}$$

where:

- $\hat{t}_{2014}^{GES}$ is the 2014 GES total crash estimate;
- t_{2014} is the 2014 true total crash count;
- $E(e_{2014}^{GES}) = 1$
- $B_{2014}^{GES} = [E(\hat{t}_{2014}^{GES}) - t_{2014}]/t_{2014}$ is the relative bias of 2014 GES total crash estimate;

Now the percent change $\hat{p}$ in the total crash estimate from 2014 to 2015 can be expressed as:

$$\hat{p} = \frac{\hat{t}_{2015}^{GES}}{\hat{t}_{2014}^{GES}} - 1$$

Notice for the same estimator of the same study variable under the same sample design, the same data collection operation and the same weighting procedure, the relative biases tend to similar, i.e., $B_{2015}^{GES} \approx B_{2014}^{GES}$, therefore,

$$E(\hat{p}) \approx \frac{t_{2015} * (1 + B_{2015}^{GES})}{t_{2014} * (1 + B_{2014}^{GES})} - 1 \approx \frac{t_{2015}}{t_{2014}} - 1$$

Similarly, the difference between the 2015 GES total crash estimate and the 2014 GES total crash estimate can be expressed as:

$$\begin{aligned}\hat{d} &= \hat{t}_{2015}^{GES} - \hat{t}_{2014}^{GES} \\ &= t_{2015} * (1 + B_{2015}^{GES}) * e_{2015}^{GES} - t_{2014} * (1 + B_{2014}^{GES}) * e_{2014}^{GES}\end{aligned}$$

Since $(1 + B_{2015}^{GES}) \approx (1 + B_{2014}^{GES})$, we have:

$$E(\hat{d}) = t_{2015} * (1 + B_{2015}^{GES}) - t_{2014} * (1 + B_{2014}^{GES}) \approx (t_{2015} - t_{2014})(1 + B_{2015}^{GES})$$

Although the difference estimate is biased by a factor $(1 + B_{2015}^{GES})$, a significance test on the difference estimate is still testing whether $(t_{2015} - t_{2014})$ is significantly different from zero as long as the relative bias $B_{2015}^{GES} \neq -100\%$.

In summary, comparisons among GES estimates or among CRSS estimates are less likely confounded by the bias. Comparison between CRSS and GES should be performed with caution.

CRSS sample selection is independent from GES sample selection. To capture this independence, a new stratification variable *STUDY* (*STUDY*=1 for GES and *STUDY*=2 for CRSS) is created. Annual samples within GES (or annual samples within CRSS in the future) are not independent samples because the same PSU and PJ samples are used for data collection. A domain (sub-population) identification variable *YEAR* is created to make year-to-year comparisons.

In the following SAS program, two years of GES data (2014 and 2015) are combined with 2016 CRSS data. First, annual crash counts by crash severity are estimated. Then pairwise comparisons are made between the annual estimates. Both analyses were implemented by SAS callable SUDAAN procedures.

In the data step, *STUDY*=1 for GES, and 2 for CRSS. Variable *YEAR* has three categories: 2014, 2015, and 2016. The two GES certainty PSUs (13 and 14) were treated as strata and the sampled PJs in those two certainty PSUs were treated as PSUs for variance estimation purpose.

The SUDAAN PROC CROSSTAB procedure produces annual parameter and variance estimates at all levels of crash severity (*CRASH_SEV*). Notice variable *STUDY* is used as an extra stratification variable so that the PSU identification variable, *PSU_VAR*, is the third variable listed in the NEST statement (*PSULEV*=3). The SUDAAN PROC DESCRIPT procedure produces pairwise comparisons between the annual estimates. The following are programs and the major output tables from those two procedures.

```
OPTIONS NOFMTERR PAGESIZE=70 LINESIZE=120;

ODS RESULTS OFF;
ODS LISTING;

LIBNAME CRSS2016 "R:\CRSS-Archive\2016\SAS File\Final\Coded
Cases for 53 PSUs - Imputed";
LIBNAME GES2015  "R:\GES\2015";
LIBNAME GES2014  "R:\GES\2014";

PROC FORMAT;
     VALUE SEVERITY 1="FATAL" 2="INJURY" 3="PDO";
     RUN;

DATA COMBINED;
     SET CRSS2016.ACCIDENT (IN=CRSS2016)
          GES2015.ACCIDENT  (IN=GES2015)
          GES2014.ACCIDENT  (IN=GES2014);
     STUDY = GES2014 + GES2015 + CRSS2016*2;
     YEAR  = CRSS2016*2016 + GES2015*2015 + GES2014*2014;
     IF (GES2015 OR GES2014) THEN DO;
          IF PSUSTRAT IN (13, 14) THEN PSU_VAR=PJ;
          ELSE PSU_VAR=PSU;
     END;
     IF MAXSEV_IM=4 THEN CRASH_SEV=1; /*FATAL CRASH*/
     ELSE IF MAXSEV_IM IN (1,2,3,5) THEN CRASH_SEV=2; /*INJURY
     CRASHES*/
     ELSE IF MAXSEV_IM IN (0,6,8) THEN CRASH_SEV=3; /*PDO
     CRASHES*/
     RUN;

PROC CROSSTAB DATA=COMBINED FILETYPE=SAS DESIGN=JACKKNIFE
NOTSORTED;
     NEST      STUDY PSUSTRAT PSU_VAR / PSULEV=3;
     WEIGHT    WEIGHT;
     CLASS          YEAR CRASH_SEV;
     TABLES    YEAR*CRASH_SEV;
     SETENV    ROWWIDTH=20 COLWIDTH=20 LABWIDTH=40;
     PRINT          NSUM="SAMSIZE" WSUM="TOTAL" SEWGT="SE TOTAL"
               / NSUMFMT=F8.0 WSUMFMT=F10.0 SEWGTFMT=F9.0;
     RFORMAT   CRASH_SEV SEVERITY.;
     RTITLE    "GES 2014, 2015 and CRSS 2016 Crash Severity
Comparison";
     RUN;

PROC DESCRIPT DATA=COMBINED FILETYPE=SAS DESIGN=JACKKNIFE
NOTSORTED TOTALS;
     NEST      STUDY PSUSTRAT PSU_VAR / PSULEV=3;
     WEIGHT    WEIGHT;
     CLASS     YEAR CRASH_SEV;
     TABLES    CRASH_SEV;
     VAR       _ONE_;
     PAIRWISE  YEAR / NAME="YEAR TO YEAR COMPARISON";
     SETENV    ROWWIDTH=20 COLWIDTH=20 LABWIDTH=40;
     PRINT     NSUM="SAMSIZE" TOTAL="DIFF" SETOTAL="DIFF STE"
               LOWTOTAL UPTOTAL
               / NSUMFMT=F10.0 TOTALFMT=F12.0 SETOTALFMT=F12.0
                 LOWTOTALFMT=F12.0 UPTOTALFMT=F12.0;
     RFORMAT   CRASH_SEV SEVERITY.;
     RUN;
```

Table 4: Crash Severity Estimates (PROC CROSSTAB)

```
Variance Estimation Method: Delete-1 Jackknife
GES 2014, 2015 and CRSS 2016 Crash Severity Comparison
by: Crash Date (Year), CRASH_SEV.

-----------------------------------------------------------------------------------------------
|                      |                  | CRASH_SEV                                         |
| Crash Date (Year)    |                  |---------------------------------------------------|
|                      |                  | Total      | FATAL      | INJURY     | PDO        |
-----------------------------------------------------------------------------------------------
|                      |                  |            |            |            |            |
| Total                | SAMSIZE          |    157,623 |      2,814 |     80,747 |     74,062 |
|                      | TOTAL            | 19,619,880 |     79,965 |  5,540,321 | 13,999,593 |
|                      | SE TOTAL         |    910,272 |      4,590 |    235,172 |    727,134 |
-----------------------------------------------------------------------------------------------
|                      |                  |            |            |            |            |
| 2014                 | SAMSIZE          |     53,030 |        895 |     27,447 |     24,688 |
|                      | TOTAL            |  6,058,524 |     24,296 |  1,647,726 |  4,386,502 |
|                      | SE TOTAL         |    404,124 |      2,572 |     94,199 |    329,222 |
-----------------------------------------------------------------------------------------------
|                      |                  |            |            |            |            |
| 2015                 | SAMSIZE          |     57,078 |        954 |     28,941 |     27,183 |
|                      | TOTAL            |  6,284,851 |     21,255 |  1,715,394 |  4,548,203 |
|                      | SE TOTAL         |    402,086 |      1,474 |     95,838 |    328,529 |
-----------------------------------------------------------------------------------------------
|                      |                  |            |            |            |            |
| 2016                 | SAMSIZE          |     47,515 |        965 |     24,359 |     22,191 |
|                      | TOTAL            |  7,276,505 |     34,415 |  2,177,201 |  5,064,889 |
|                      | SE TOTAL         |    438,260 |      3,141 |    142,291 |    322,942 |
-----------------------------------------------------------------------------------------------
```

The 2014, 2015 and 2016 FARS fatal crash counts are 30,056, 32,166, and 34,439 respectively (see NHTSA’s “Traffic Safety Facts 2015” and “Traffic Safety Facts 2016”). The 2014 and 2015 GES fatal crash estimates (24,296 with standard error 2,572 and 21,255 with standard error 1,474) are significantly lower than their FARS counterparts. The difference between the 2016 CRSS fatal estimate and the 2015 GES fatal estimate (34,415 – 21,255 = 13,160) contains the real difference and the bias.

Table 5: Pairwise Comparisons by Crash Severity (PROC DESCRIPT)

```
GES 2014, 2015 and CRSS 2016 Crash Severity Comparison
by: Variable, CRASH_SEV, Contrast.

---------------------------------------------------------------------------------------
|                       |                   | Contrast                                 |
| CRASH_SEV             |                   |------------------------------------------|
|                       |                   | YEAR TO YEAR | YEAR TO YEAR | YEAR TO YEAR |
|                       |                   | COMPARISON:  | COMPARISON:  | COMPARISON:  |
|                       |                   | (2014,2015)  | (2014,2016)  | (2015,2016)  |
---------------------------------------------------------------------------------------
|                       |                   |              |              |              |
| Total                 | SAMSIZE           |      110,108 |      100,545 |      104,593 |
|                       | DIFF              |     -226,327 |   -1,217,982 |     -991,654 |
|                       | DIFF STE          |      115,987 |      596,144 |      594,764 |
|                       | Lower 95% Limit   |              |              |              |
|                       |  Cntrst Total     |     -456,901 |   -2,403,076 |   -2,174,006 |
|                       | Upper 95% Limit   |              |              |              |
|                       |  Cntrst Total     |        4,246 |      -32,887 |      190,698 |
---------------------------------------------------------------------------------------
|                       |                   |              |              |              |
| FATAL                 | SAMSIZE           |        1,849 |        1,860 |        1,919 |
|                       | DIFF              |        3,041 |      -10,119 |      -13,161 |
|                       | DIFF STE          |        2,526 |        4,060 |        3,470 |
|                       | Lower 95% Limit   |              |              |              |
|                       |  Cntrst Total     |       -1,979 |      -18,190 |      -20,058 |
|                       | Upper 95% Limit   |              |              |              |
|                       |  Cntrst Total     |        8,062 |       -2,049 |       -6,263 |
---------------------------------------------------------------------------------------
|                       |                   |              |              |              |
| INJURY                | SAMSIZE           |       56,388 |       51,806 |       53,300 |
|                       | DIFF              |      -67,668 |     -529,475 |     -461,807 |
|                       | DIFF STE          |       32,521 |      170,647 |      171,557 |
|                       | Lower 95% Limit   |              |              |              |
|                       |  Cntrst Total     |     -132,317 |     -868,709 |     -802,850 |
|                       | Upper 95% Limit   |              |              |              |
|                       |  Cntrst Total     |       -3,018 |     -190,241 |     -120,764 |
---------------------------------------------------------------------------------------
|                       |                   |              |              |              |
| PDO                   | SAMSIZE           |       51,871 |       46,879 |       49,374 |
|                       | DIFF              |     -161,701 |     -678,387 |     -516,687 |
|                       | DIFF STE          |       90,587 |      461,172 |      460,677 |
|                       | Lower 95% Limit   |              |              |              |
|                       |  Cntrst Total     |     -341,782 |   -1,595,166 |   -1,432,482 |
|                       | Upper 95% Limit   |              |              |              |
|                       |  Cntrst Total     |       18,380 |      238,391 |      399,109 |
---------------------------------------------------------------------------------------
```

The standard errors for the differences (DIFF STE) between GES 2014 and GES 2015 are much smaller than those between GES and CRSS annual estimates (GES 2014 versus CRSS 2016 or GES 2015 versus CRSS 2016) because GES annual estimates are positively correlated while the GES estimate is independent from the CRSS estimate. It should be noted that the difference between CRSS estimates and GES estimates may be confounded by the potential GES bias.

5.3. Example 3: Composite Estimates

A composite estimate refers to a function of estimates made from different samples. FARS is an annual census survey of all fatal crashes. FARS data collection is independent from GES or CRSS data collection. FARS's target population – fatal crashes - is a sub-population of GES or CRSS target population. FARS finite population parameter estimates do not have design variance because all fatal crashes were observed. Therefore, FARS data can provide very good estimates for the fatal domain (sub-population). Because of this, it is sensible to make a composite estimate using estimates from both FARS and CRSS as in the following example:

- Make an estimate $\hat{\theta}_{FARS}$ for total fatal crash count from FARS.
- Make an estimate $\hat{\theta}_{CRSS}$ for total injury crash count from CRSS non-fatal domain.
- Create a composite estimate for total fatal and injury count: $\hat{\theta}_C = \hat{\theta}_{FARS} + \hat{\theta}_{CRSS}$.

Specifically, to estimate the total number of fatal and injury crashes in 2016, first calculate the total number of fatal crashes from 2016 FARS: $\hat{\theta}_{FARS}$ =34,439. Then from Example 2 above, estimate the number of injury crashes from 2016 CRSS: $\hat{\theta}_{CRSS}$=2,177,201. Finally, the total number of fatal and injury crashes in 2016 is estimated by composite estimate:

$$\hat{\theta}_C = 34{,}439 + 2{,}177{,}201 \ = \ 2{,}211{,}640$$

The design variance of $\hat{\theta}_C$ is:

$$\begin{aligned} Var_d(\hat{\theta}_C) &= Var_d(\hat{\theta}_{FARS}) + Var_d(\hat{\theta}_{CRSS}) \\ &= Var_d(\hat{\theta}_{CRSS}) = 142{,}291^2 \end{aligned}$$

Covariance of $\hat{\theta}_{FARS}$ and $\hat{\theta}_{CRSS}$ is zero because $\hat{\theta}_{FARS}$ and $\hat{\theta}_{CRSS}$ are independent. $Var_d(\hat{\theta}_{FARS}) = 0$ because FARS is a census, and $Var_d(\hat{\theta}_{CRSS}) = 142{,}291^2$ from Table 4.

It should be noted that the domain estimate $\hat{\theta}_{CRSS}$ (total injury crash count) should be calculated from the full CRSS sample as in Example 2. Sub-setting the full sample for domain estimation may produce a biased variance estimate (see Graubard et al., 1996).

Another approach is combining FARS data with the same year CRSS data and then making estimates from the combined data. However, because the same year FARS and CRSS data have overlapping coverages for the fatal crashes, dual frame estimation methods should be used to analyze the combined data. See Lohr and Rao (2000) and Mecatti (2007) for more information about dual frame estimation.

5.4 Example 4: Domain Estimates

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 sub-setting the full sample for domain estimation.

In SAS PROC SURVEY procedures, domains are specified by the variables listed in the TABLES and/or the DOMAIN statement. The SAS BY statement sub-sets 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 percentage of alcohol-involved non-fatal crashes. To this end, we classify all crashes into two domains: fatal and non-fatal crashes. The domains were defined by variable *FATAL* in the DOMAIN statement.

```
PROC FORMAT;
     VALUE FATAL 1="FATAL" 0="NON-FATAL";
     RUN;

DATA CRSS2016;
     SET COMBINED;
     IF YEAR=2016;
     FATAL=(CRASH_SEV=1);    /*FATAL=1 IF FATAL, 0 OTHERWISE*/
     ALCOHOL=(ALCHL_IM=1); /*ALCOHOL=1 IF ALCOHOL INVOLVED*/
     RUN;
```

```
PROC SURVEYMEANS DATA=CRSS2016 VARMETHOD=JK MEAN SUM CLM;
     STRATA     PSUSTRAT;
     CLUSTER    PSU_VAR;
     WEIGHT     WEIGHT;
     VAR        ALCOHOL;
     DOMAIN     FATAL;
     FORMAT     FATAL FATAL.;
     RUN;
```

Table 6: SAS PROC SURVEYMEANS domain estimates

```
                                   Domain Analysis: FATAL

                                                                Std Error
 FATAL      Variable  Label                           Mean        of Mean       95% CL for Mean
------------------------------------------------------------------------------------------------
 NON-FATAL  ALCOHOL   Alcohol Involved In Crash   0.043592       0.002480   0.03852813 0.04865657
 FATAL      ALCOHOL   Alcohol Involved In Crash   0.209324       0.017185   0.17422747 0.24442118
------------------------------------------------------------------------------------------------
```

The following SAS-callable SUDAAN program also estimates the percentage of alcohol involved crashes by two domains: fatal and non-fatal crashes. The domains were defined by variable *FATAL* in the TABLE statement.

```
PROC DESCRIPT DATA=CRSS2016 DESIGN=JACKKNIFE NOTSORTED;
     NEST       PSUSTRAT PSU_VAR;
     WEIGHT     WEIGHT;
     CLASS      FATAL;
     TABLES     FATAL;
     VAR        ALCOHOL;
     SETENV     ROWWIDTH=15 COLWIDTH=15 LABWIDTH=15;
     PRINT      NSUM="SAMSIZE" WSUM="POPSIZE" MEAN="MEAN"
                SEMEAN="MEAN SE" LOWMEAN UPMEAN
                / NSUMFMT=F8.0 WSUMFMT=F10.0 MEANFMT=F6.4
                SEMEANFMT=F6.4 LOWMEANFMT=F6.4 UPMEANFMT=F6.4;
     RUN;
```

Table 7: SUDAAN domain estimates

```
-----------------------------------------------------------------------------
|                 |                  | FATAL                                  |
| Variable        |                  |----------------------------------------|
|                 |                  | Total        | 0            | 1            |
-----------------------------------------------------------------------------
|                 |                  |              |              |              |
| Alcohol         | SAMSIZE          |       47,515 |       46,550 |          965 |
| Involved In     | POPSIZE          |    7,276,505 |    7,242,090 |       34,415 |
| Crash           | MEAN             |       0.0444 |       0.0436 |       0.2093 |
|                 | MEAN SE          |       0.0025 |       0.0025 |       0.0172 |
|                 | Lower 95% Limit  |              |              |              |
|                 |  Mean            |       0.0392 |       0.0385 |       0.1742 |
|                 | Upper 95% Limit  |              |              |              |
|                 |  Mean            |       0.0495 |       0.0487 |       0.2444 |
-----------------------------------------------------------------------------
```

6. Frequently Asked Questions

1. **What is CRSS?**

   A. The Crash Report Sampling System is NHTSA's new national probability-based crash sampling system designed to replace the General Estimates System.

2. **What data does CRSS collect and what does it represent?**

   A. Like GES, CRSS samples police crash reports and codes the information into a data file. The CRSS data, when used with the accompanying weights, are nationally representative of all police-reported motor vehicle traffic crashes where the first harmful event occurred on a public trafficway.

3. **When did NHTSA transition from GES to CRSS?**

   A. 2015 was the last year of data collection through GES. CRSS was designed and implemented over a multi-year effort and started collecting data in January 2016.

4. **Why did NHTSA transition from GES to CRSS?**

   A. The Congress directed and provided funds to NHTSA to modernize its data collection system. Since the GES had used the same data collection sites since 1988, the existing GES police jurisdiction samples and weights became outdated as the PJ population changed. In addition, GES produced biased fatal crash estimates. Given the shifts in population and vehicle fleet, and the changing analytic needs of the safety community, NHTSA modernized its crash data collection system.

5. **How is CRSS different from GES in terms of its sample design?**

   A. The following are some major differences in sample designs of CRSS and GES:

   - Independent sample: the CRSS sample design is independent from the GES or any other NHTSA's surveys, including NHTSA's new Crash Investigation Sampling System that replaces the NASS Crashworthiness Data System (CDS). 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 - CRSS and CISS.
   - Different formation of PSUs: In both CRSS and GES, a PSU is either a county or a group of counties. In CRSS, the nation was partitioned into 707 PSUs, while in GES 1195 PSUs were formed. CRSS's average PSU size is bigger

than GES. This resulted in more equal weights in CRSS. In addition, a new composite PSU measure of size variable using the various estimated crash counts by the new PAR strata was used in CRSS.

- Finer PSU stratification: The 60 GES PSUs were selected from 12 PSU strata formed by census region and urbanicity type. The 60 CRSS PSUs were selected from 25 PSU strata formed by census region, urbanicity type, vehicle miles traveled, total number of crashes, total truck miles traveled, and road miles.
- Scalable PSU sample: the CRSS 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 CRSS 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: At CRSS PAR sample selection stage, PAR strata were revised based on data needs to oversample the following analysis domains:
    - Crashes involving killed or injured pedestrian;
    - Crashes involving killed or injured motorcycle occupant;
    - Crashes involving killed or injured occupant of late model year passenger vehicle; and
    - Crashes involving killed or severely injured occupant of non-late model year passenger vehicle.
- Optimized sample allocation: CRSS PSU, PJ and PAR sample allocation was optimized by minimizing the variance of a simplified variance estimator subject to fixed cost.

- Flexible system to allow mid-year changes: CRSS allows for mid-year changes to sampling parameters such as PAR sampling intervals, PJ sample changes etc. to cope with unanticipated changes. GES sample parameters could not be changed in the middle of the year.

6. **How is CRSS similar to GES in sample design?**

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

   - The CRSS target population is the same as the GES, i.e., all police-reported crashes of motor vehicles occurring on a public trafficway.
   - Both CRSS and GES collect information mainly from police crash reports.
   - Both CRSS and GES have a three-stage sample design: PSU, PJ, and PAR sample selection.
   - In both surveys, PSUs and PJs have selection probabilities proportional to their measure of sizes.
   - Both surveys' PAR samples were selected using systematic sampling.
   - Both surveys tried to achieve equal-weights within PAR stratum.

7. **How do the CRSS analysis files (data sets) differ from the GES?**

   A. The CRSS analysis file is almost the same as the GES analysis file. They have the same variables with the same names except CRSS no longer codes the land use variable. A new PSU stratification variable *PSU_VAR* was created in CRSS for variance estimation purposes (see Example 1).

8. **Is the difference in total crashes (or other metric of choice) between the 2015 GES and the 2016 CRSS estimates due to the design change or an actual increase?**

   A. The difference between the 2015 GES total crash estimate and 2016 CRSS total crash estimate may result from the following:

   - The difference between the true 2015 total crash count and the true 2016 total crash count;
   - The sampling errors of 2015 GES total crash estimate and 2016 CRSS total crash estimate;
   - The potential bias of 2015 GES total crash estimate.

For example, in the past we have observed GES fatal crash underestimation. When a CRSS estimate is compared to a biased GES estimate, the observed difference is confounded by the GES bias. However, different variables suffered different degrees of bias. The comparisons among GES estimates are less likely affected by the bias. See Example 2 for more discussion on this.

9. **Is the difference between the 2015 GES total crash estimate and the 2016 CRSS total crash estimate statistically significant?**

   A. No. The difference between the 2016 CRSS total crash estimate and the 2015 GES estimate is not significantly different from zero due to the large variance of the difference estimate. It should be noted this only means given the data we have, we do not have evidence to reject the null hypothesis that the two years have the same total number of crashes. This does not necessarily mean the null hypothesis is true. See Example 2 for more detailed information about this test and computer codes.

10. **Can we combine FARS and CRSS data for estimates and analysis?**

    A. Yes. Below are two options of combining FARS and CRSS data:

    - Combining estimates: Estimates are made separately from FARS data file and CRSS sample file. Then a composite estimate is created by combining the estimates (Example 3, Chapter 5).
    - Combining data sets: The FARS data file and the CRSS sample file are combined first. Weights are adjusted for the overlapping subpopulations. Then an estimate is made from the combined data set. This is also called dual frame estimation. See Lohr and Rao (2000) and Mecatti (2007) for more information about dual frame estimation.

11. **How should data users compute the variance for CRSS estimates?**

    A. The CRSS 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 PROC SURVEY procedures or SUDAAN procedures should be used to make estimates from CRSS 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. See Chapters 4 and 5 for some basic concept of complex survey data analysis, and SAS and SUDAAN sample programs on how to estimate the variances for CRSS estimates.

12. 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 CRSS sample. Some examples of such softwarse: SAS PROC SURVEY procedures, SUDAAN, R survey package, and STATA. See Chapter 5 for specific examples of programming techniques to achieve variance estimation.

13. Are there issues of bias with GES estimates?

A. Yes. As we have seen in the past, GES fatal crash estimates and fatality estimates were significantly lower than the corresponding FARS counts. Therefore, NHTSA has been using FARS data for fatal crash counts. Different variables may suffer different degrees of bias. But trends within GES are less likely to be affected by the bias.

14. How are the CRSS estimates validated?

A. Broadly, NHTSA used known population parameters such as fatal crashes and census population estimates to assess the performance of PSU weights. In addition, NHTSA collected PSU level total crash counts to validate the PJ and PAR weights. These evaluations established the reliability of the weights and the results will be documented in the upcoming CRSS sample design and weighting report.

15. If GES underestimated fatal crashes, how has this issue been addressed in CRSS? Will it still be necessary to replace CRSS fatal crashes with FARS data, or would CRSS be able to stand alone for all analyses of fatal/non-fatal crashes?

A. Based on the evaluation of the 2016 CRSS estimates, the CRSS does not underestimate fatal crash counts. CRSS can be used to make fatal estimates, but using FARS data is recommended as FARS is a nationwide census of all fatal crashes. NHTSA will continue to use FARS for fatality counts and CRSS for injury and PDO related estimates.

16. **Is there anything different need to do with CRSS data in producing estimates for very small sample size?**
    A. This is a small area estimation problem. The problem associated with a small sample size does not change from GES to CRSS. See Rao and Molina 2015 for more details on small area estimation.

17. **How are missing data addressed in CRSS?**
    A. As in GES, key CRSS variables with missing information are imputed using the sequential regression multivariate imputation procedure. The details of this procedure will be illustrated in the CRSS Analytical User's Manual.

18. **Are there significant differences between the GES imputation and the CRSS imputation?**
    A. No.

7. References

Graubard, B. I., & Korn, E. L. (1996). Survey inference for subpopulations. *American Journal of Epidemiology, 144*(1), pp 102-106.

Hartley, H. O., & Sielken, R. L. (1975). A "super-population viewpoint" for finite population sampling. *Biometrics, 31*(2), pp 411-422.

Lohr, S., and Rao, J. N. K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association, 95*(449), pp. 271–280.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys, *Survey Methodology, 33*(2), pp. 151-157.

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

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

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 Report Sampling System: Sample Design and Estimation. Washington, DC: National Highway Traffic Safety Administration.

DOT HS 812 509
March 2018

U.S. Department
of Transportation

National Highway
Traffic Safety
Administration

NHTSA