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

NHTSA
www.nhtsa.gov

DOT HS 811 327

May 2010

Sampling and Estimation Methodologies of CDS

DISCLAIMER

This publication is distributed by the U.S. Department of Transportation, National Highway Traffic Safety Administration (NHTSA), 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.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. DOT HS 811 327	2. Government Accession No.	3. Recipients Catalog No.
4. Title and Subtitle Sampling and Estimation Methodologies of CDS		5. Report Date May 2010
		6. Performing Organization Code NHTSA/NVS-421
7. Author(s) Charles Fleming (NCSA)		8. Performing Organization Report No.
9. Performing Organization Name and Address National Highway Traffic Safety Administration 1200 New Jersey Avenue SE Washington, DC 20590		10. Work Unit No. 11. Contract or Grant No.
12. Sponsoring Agency Name and Address National Center for Statistics and Analysis (NCSA) National Highway Traffic Safety Administration 1200 New Jersey Avenue SE, Washington, DC 20590		13. Type of Report and Period Covered Technical Report 14. Sponsoring Agency Code NHTSA
15. Supplementary Notes Thanks for help and input from Nancy Bondy, Ruby Li, and Donna Glassbrenner.		
16. Abstract Based primarily on two computer programs and expert knowledge, this document describes the method of sampling and the method of estimation which are used in the Crashworthiness Data System.		
17. Key Words Sampling, estimation, National Automotive Sampling System, Crashworthiness Data System	18. Distribution Statement: Document is available to the public from the National Technical Service www.ntis.gov	

19. Security Classification (of this report) UNCLASSIFIED	20. Security Classification (of this page) UNCLASSIFIED	21. No. of Pages 27	22. Price
Form DOT F 1700.7 (8-72)	Reproduction of completed page authorized		

Table of Contents

List of Figures

List of Tables

1 Introduction

The origins of the Crashworthiness Data System can be traced to the National Accident Sampling System which was implemented for operational use in 1979. Important documents which describe the design of the original system can be found in [1, 2, 3, 6, 7, 9, 11, 12, 13, 14, 15, 16, 18, 19]. Later, its name was changed to the National Automotive Sampling System (NASS). In 1988, NASS was divided into two parts: the General Estimates System (GES) and the Crashworthiness Data System (CDS). The CDS is one of the crash databases produced by NHTSA. A general description of CDS can be found in [8].

This report describes the sampling and estimation methodologies of CDS as they are followed in the operational program. The primary sources of information upon which this report is based are four documents. They are [4], the two SAS computer programs, `PSUWGHT.SAS` and `CDSWGT.SAS`, and sampling worksheets for manually drawing a sample of police jurisdictions (PJ). The purpose of these worksheets is to validate the computerized sampling algorithm, in order to make sure that the cases are selected the same way whether manually or by computer program. This computer program which is written in DELPHI performs the same task automatically as that of the sampling worksheets. The output of the DELPHI computer program has been validated by the sampling manager of CDS for accuracy.

The formulas which appear in this report were inferred from the computer programs, and their accuracy was confirmed by comparing computations based on these formulas with the output of the computer programs. To make it easier for the reader to associate the formulas with the logic of the computer programs, names of the variables which appear in the SAS programs are retained in the formulas. The correspondence between the names of SAS variables and the variables of the final analytical files is always maintained except in three cases. The three variables of the computer programs which are not the same names which appear in the analytical files are listed in Table 1.

Table 1: Computer Program Variable Names in Analytical Files

Variable Name in Computer Program	Variable Name in Analytical Files
PSUGRP	PSUSTRAT
RATWGHT	RATWGT
STRATA	STRATIF

This report is divided into two main parts. The first part describes the method of sampling which is used in CDS. Afterwards, the second part describes the method of estimation. The same terminology as well as the definitions of the population of interest and of the stratification which are published in [10] are followed in this report. Three levels of stratification appear in CDS. The stratification of the PSU's constitutes the first level. It is followed by the stratification of the PJ's, and then the stratification of the police accident reports (PAR) follows next. To emphasize the importance of the stratification of the PAR's, the definitions of the strata are reproduced in Appendix III.

2 Sampling

2.1 First Stage: Primary Sampling Units

The collection of primary sampling units (PSU) forms an area frame of the United States. There are 1,195 PSU's [4] and each one is assigned to one of twelve strata [17] which are defined according to geographical and demographic characteristics. The definitions of the strata and a detailed account of their determinations can be found in [17]. From the collection of PSU's, 24 of them were selected in 1991 such that two PSU's are selected from each of twelve strata. In a demonstration of the versatility of the design of CDS to changing conditions, three more PSU's were added in 2002 for a special research study until 2008 when the original 24 PSU's again comprise the sample of PSU's. For this special study, three PSU's were added from 2002 through 2004 to the sampling frame and the sampling was restricted to include only late model year motor vehicles. From 2005 through 2007, the CDS continued to include the three extra PSU's, but they included both late model year and non-late model year vehicles. The currently used PSU's for sampling have been used ever since CDS began, although a re-definition of their sampling weights from the 1979 values did occur in 1991 [4].

The method of calculating the PSU weights is explained in [4], and can be summarized as

1. Counts of fatal and injury crashes were obtained for each of the 1,195 PSU's.
2. For all PSU's within each PSU stratum, the number of fatal and injury crashes were summed.
3. The probability of a PSU being selected was determined by dividing the number of fatal and injury crashes occurring within a PSU by the total number of fatal and injury crashes occurring in all of the PSU's within that stratum. The weight of the PSU is simply the inverse of the probability of the PSU being selected.

That the number of PSU's should change to accommodate a special study shows the flexibility of the CDS to changes. For example, in 1996, the Agency had to accommodate a sudden increase in the number of cases involving deaths and serious injuries which were associated with air bags. In response to this surge of cases, more PJ's were added, in order to obtain more crashes involving late model year vehicles.

Groups of PSU's (PSUGRP) have been created according to Table 2 for use in the computer program for calculating the final published weights. There is a one-to-one correspondence between PSUGRP and PSUSTRAT, although the numbering scheme is different between them. Appendix I shows the correspondence between PSUGRP and PSUSTRAT. Nonetheless, the final weights which are denoted by RATWGHT in the SAS computer programs and by RATWGT in the analytical files are eventually associated with only a PSU stratum rather than to a specific PSU. Besides having two PSU's belonging to a PSU stratum, the strata of the PAR's are grouped into categories called COLSTRAT according to the assignments given in Table 3 such that the final published weights, RATWGT, correspond to a particular PSUSTRAT and to a particular COLSTRAT.

Table 2: PSUGRP

PSUGRP	PSU
1	3,6
2	72,74
3	41,49
4	79,82
5	5,8
6	12,73
7	9,45
8	75,81
9	2,4
10	11,13
11	43,48
12	76,78

Table 3: COLSTRAT

COLSTRAT	Strata
AB	$A \cup B$
CJ	$C \cup J$
DK	$D \cup K$
E	E
F	F
G	G
H	H

2.2 Second Stage: Police Jurisdictions

Within each PSU, there are areas called police jurisdictions (PJ). Within a police jurisdiction there is an administrative center where police accident reports (PAR) are assembled and stored. It should be noted that the preferred reference to the abbreviation, PAR, is to call it a police crash report. Because PAR's may be in an electronic format, typewritten paper copy, or hand written paper copy and since they may be issued at any time of the day, the logistical demands of drawing a sample of PAR's require special consideration.

The method of selecting a police jurisdiction is based on probability proportional to size (PPS) sampling though exceptions are sometimes made. Even though, according to the original design of NASS, police jurisdictions are supposed to be re-selected periodically at random, the same police jurisdictions like the PSU's have been used since 1995 .

To take into account varying sizes of the PJ's, police jurisdictions are classified by strata in such a way so as to create groups of PJ's which have equal number of fatal and injury crashes within a PSU. These groups are in essence strata which are determined by the number of instances in which at least one death, at least one incapacitating injury, or at least one non-incapacitating injury had occurred. This number is abbreviated by KAB. It serves as the measure of size of a PJ when selecting a police jurisdiction within one of these PJ strata according to the method of probability proportional to size. The current CDS sampling uses 1992 KAB data. The number of PJ strata varies according to PSU, and it depends on the goal to equalize the number of cases across PJ's within a PJ stratum. The method which is followed begins by sorting the PJ's of a PSU by the number of fatal and injury crashes in descending order of magnitude. Then the PJ's are grouped into strata of similar size. Usually, the largest strata contain only one PJ which, in turn, is selected with certainty.

2.2.1 Method of Selecting Police Jurisdictions

In explaining the method of selecting a police jurisdiction, PJ_{il}, from PSU_i, it is useful to define certain variables which closely mimic the variables which appear in the computer programs. We will use concepts which are associated with the method of probability proportional to size sampling in which the value KAB is the measure of size of a police jurisdiction and only one PJ is selected from a PJ stratum.

Definition 1. *KAB is an abbreviation for killed, incapacitating injury, and non-incapacitating injury as defined by the KABCO [5] system of classifying the severity of injury. Let KAB_{il} be the number of cases in which at least one death or at least one incapacitating injury or at least one non-incapacitating injury occurred in PSU_i and in PJ_{il}.*

Table 4 presents a list of PJ's and the associated KAB's for a hypothetical PSU_i. We will refer to Table 4 when describing the method of selecting PJ's.

Table 4: Calculation of $PJWGHT$

PJ		PJSTRAT				Random Start			
l	KAB_l	m	s	$cumsum_{s\le n_m}(KAB_{ms})$	$RAND_m$	$RAND_m \sum_{s=1}^{n_m} KAB_{ms}$	Select	λ_m	$PJWGHT_l$
1	85	1	1	85	1	85	1	1	1
2	81	2	1	81	1	81	1	2	1
3	73	3	1	73	1	73	1	3	1
4	67	4	1	67	.016	.016(133)=2	1	4	$\frac{133}{67} = 1.99$
5	66	4	2	133			0		
6	65	5	1	65	.504	.504(118)=59	1	6	$\frac{118}{65} = 1.82$
7	53	5	2	118			0		
8	47	6	1	47	.935	.935(89)=83	0		
9	42	6	2	89			1	9	$\frac{89}{42} = 2.12$
10	35	7	1	35	.258	.258(116)=30	1	10	$\frac{116}{35} = 3.31$
11	31	7	2	66			0		
12	20	7	3	86			0		
13	15	7	4	101			0		
14	15	7	5	116			0		
15	15	8	1	15	.368	.368(55)=20	0		
16	14	8	2	29			1	16	$\frac{55}{14} = 3.93$
17	12	8	3	41			0		
18	10	8	4	51			0		
19	4	8	5	55			0		
20	0	8	6	55			0		

Definition 2. *Let l designate a PJ of a PSU, so that PJ_{il} is the l^{th} PJ in the i^{th} PSU, and let m designate the PJ stratum.*

Definition 3. $PJSTRAT_{im} = \{PJ \in PJ\ stratum$ *m of* $PSU_i\}$*, and let* n_{im} *be the size of* $PJSTRAT_{im}$*.*

$PJSTRAT_{im}$ is a collection of all PJ's which are assigned to PJ stratum m. For example, for PJ stratum 8 as given in Table 4 of PSU_i, $PJSTRAT_{i8} = \{15, 16, 17, 18, 19, 20\}$. A catalogue of the number of PJ strata for each PSU is given in Appendix I.

Definition 4. *Let $RAND_{im}$ be the random number which is generated by a computer software package for a $U(0,1)$ probability distribution for PSU_i and $PJSTRAT_{im}$.*

Every element of a PJ stratum will be assigned the same random number. That is, every PJ stratum is assigned a unique random number and every element of that stratum is associated with that assigned random number.

Example 1 presented below illustrates the process of selecting a PJ as well as of making the mathematical description of the process more transparent. A key concept which is essential for the process of selecting an element for a sample when following the method of probability proportional to size is the cumulative sum.

Definition 5. *The cumulative sum of KAB's up to and including the KAB of the Λ^{th} PJ is defined to be* $\underset{s\leq\Lambda}{cumsum}(KAB_{is}) = \sum_{s=1}^{\Lambda} KAB_{is}$.

We see an example of cumsum in the fifth column of Table 4. Let us consider $PJSTRAT_{i8}$. The cumsum for this PJ stratum is $\{15, 29, 41, 51, 55, 55\}$. Each term of the cumsum is the sum of the previous values of KAB's including itself. For instance, the third cumsum is: 15+14+12=41. That PJ which is ultimately selected for the sample will be denoted by a special subscript, λ_m. More formally, for the purpose of selecting PJ's, let the subscript of that PJ of $PJSTRAT_{im}$ which is selected for sampling be denoted by λ_m. Its precise definition is given by

Definition 6. $\lambda_m = min\left\{\lambda | \sum_{s=1}^{\lambda} KAB_{is} \geq RAND_{im} \sum_{s=1}^{n_{im}} KAB_{is}\right\}$

where n_{im} is the number of PJ's which are contained in $PJSTRAT_{im}$. In other words, λ_m identifies that PJ of $PJSTRAT_{im}$ which is selected for the sample; therefore, $PJ_{i\lambda_m}$ is that PJ which is selected from $PJSTRAT_{im}$. We will use the cumulative summation to define a step function which lies at the basis of selecting the PJ's by means of PPS sampling. In terms of cumulative summations, Definition 6 can be written as

$$\lambda_m = min\left\{\lambda | \underset{s\leq\lambda}{cumsum}(KAB_{is}) \geq RAND_{im} \sum_{s=1}^{n_{im}} KAB_{is}\right\}$$

Denote the right hand side of the inequality by X, in order to help us to see how λ_m is determined.

$$f(X) = min\left\{\lambda | \underset{s\leq\lambda}{cumsum}(KAB_{is}) \geq X\right\} \tag{1}$$

A graphical depiction of f(X) appears in Figure 1. It is a step function in which the steps occur at each level, Λ, of $\underset{s\leq\Lambda}{cumsum}(KAB_{is})$ where the left end of a level line is empty and the corresponding right end is closed. Only the initial four of several levels are shown in Figure 1. For a given X, like the one drawn in Figure 1, equation (1) says that the chosen element for the sample is determined by finding all cumsum's of KAB which either equal X or exceed X and from that list find the PJ with the minimum cumsum. In reference to Figure 1, the cumsum's which are bigger than or equal to X are the 3^{rd}, 4^{th}, 5^{th}, 6^{th}, etc. The smallest one is the third cumsum; therefore, of the PJ's, the third PJ of

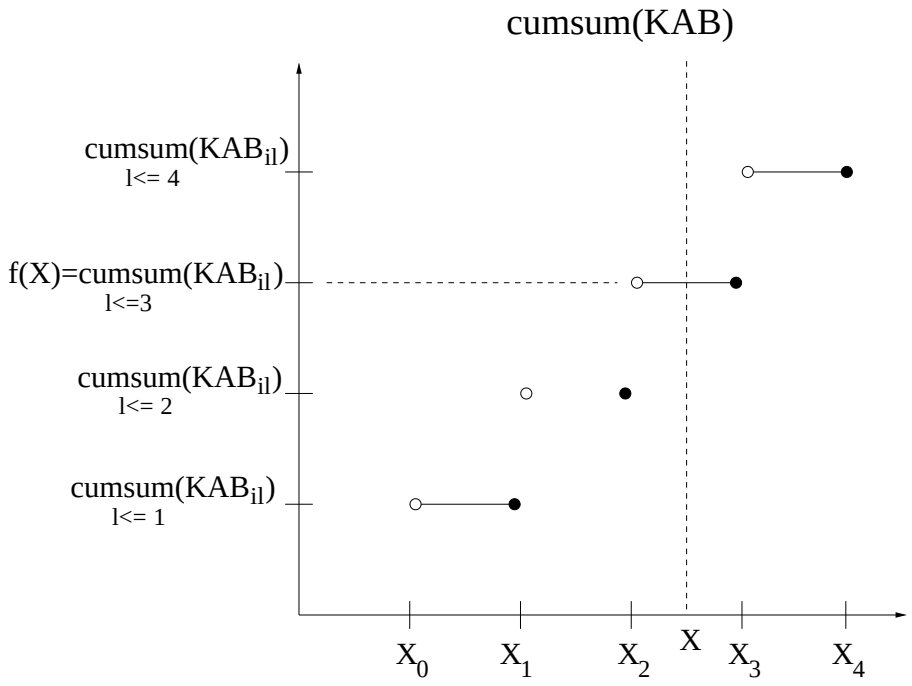

Figure 1: cumsum Function

the PJ stratum is selected. As X changes, so might f(X), and since a different $RAND_{im}$ will produce a different X, we might select a different PJ, hence we see the origins of the random sampling of the PJ's.

Definition 6 describes the basic idea of selecting a PJ:

1. List all PJ's of a PSU in descending order according to KAB's.
2. Define the boundaries of the PJ strata such that the sum of the KAB's are approximately the same across the strata and such that it is feasible for the researchers to accomplish their assignments.
3. Order the PJ's according to their KAB's within a PJ stratum.
4. Determine those PJ's which will be selected with probability 1.
5. Given a PJ stratum, list all cumsum's of KAB's which equal or exceed a random start.
6. Pick that PJ which corresponds to the smallest cumsum of that list.

Referring to Table 4 and to $PJSTRAT_{i8}$, we see that $RAND_{i8} = .368$. The random start is $.368(55) = 20$. All PJ's with cumsum's exceeding 20 are $\{16, 17, 18, 19, 20\}$. The PJ with the smallest cumsum is PJ_{i16}. Because it is the second PJ in $PJSTRAT_{i8}$, $\lambda_8 = 16$. This second PJ is the one selected for the sample.

Another example to consider is $PSTRAT_{i6}$. The cumsum is $\{47, 89\}$. The random number which is produced from a computer program is .935. The random start is .935(89)=83. The cumsum which

exceeds 83 is $\{89\}$. The smallest cumsum is 89, and it corresponds to PJ_{i9} which is the second PJ in $PSTRAT_{i6}$, hence, $\lambda_6 = 9$.

In summary, $PJ_{i\lambda_m}$ is the selected police jurisdiction of $PJSTRAT_{il}$ when it occurs that

$$\lambda_m = min \quad \lambda | \sum_{s=1}^{\lambda} KAB_{is} \geq RAND_{im} \sum_{s=1}^{n_{im}} KAB_{is}$$

The probability of the event of selecting $PJ_{i\lambda_m}$ is

$$P(PJ_{i\lambda_m}) = \frac{KAB_{i\lambda_m}}{\sum_{s=1}^{n_{im}} KAB_{is}}$$

Its reciprocal is called $PJWGHT_{i\lambda_m}$, and it is given by

$$PJWGHT_{i\lambda_m} = \frac{\sum_{s=1}^{n_{im}} KAB_{is}}{KAB_{i\lambda_m}} \tag{2}$$

Example 1. *The example given in Table 4 illustrates the method which is employed in CDS to select a police jurisdiction.*

In this example, we notice that in PJ strata 1, 2, and 3 there is only one PJ. If a total KAB count of a PJ exceeds 70, then it is selected with certainty.

The selection of an element for a sample based on PPS sampling depends on the order of the PJ's. A different order of the PJ's will produce different sets of cumsum's. When selecting PJ's, the PJ's are ordered with respect to KAB in descending order such that the PJ with the largest KAB is listed first.

2.3 Final Stage: Police Crash Reports

The statistical enumerators who are called researchers enter the PAR's from the sampled PJ's into a computer program for case selection. The researchers conduct vehicle and scene inspections within four days of the case being selected. They conduct interviews, draw the scene diagram, code the case, and obtain medical records. When determining the number of PJ's to be selected in a PSU, consideration is given to the area of the PJ, the number of jurisdictions which investigate motor vehicle traffic crashes, the number of those which involve an injury or death, the estimated number of cases which should be drawn from a PSU, and the distance between PJ's.

Sampling of PAR's is performed weekly in all 171 police jurisdictions which are located in 24 PSU's. A researcher, who is a trained statistical enumerator, once a week examines the PAR's which apply to the police jurisdiction to which he is assigned. The researcher will classify each PAR as to stratum, only if the PAR qualifies for admission into the sampling frame. Elements of the CDS sampling frame of any police jurisdiction must comply with the same criteria for admission to the sampling frame as the criteria which defines the population [10]. The weekly drawing of a PAR from a police jurisdiction sampling frame is the last stage of sampling.

The list of PAR's which is constructed each week at a police jurisdiction is divided into ten strata. These ten strata are: A, B, C, D, E, F, G, H, J, and K. The definitions of them are given in Appendix III. The last two strata were added to the original strata in 1991.

In the late 1980's, the Agency required that the number of serious injury crashes which are selected for the sample be increased. A group in the National Center for Statistics and Analysis (NCSA) considered many options to achieve the desired sampling size. Each option was analyzed by such factors as the number of potential cases by severity of injury both in terms of weighted and unweighted values and the feasibility and ease of implementation. It was determined that adding two strata, J and K, designated for hospitalization would significantly increase the number of cases of serious injuries while maintaining the weighting requirement. A pilot study was conducted for three months in 1990 to discover whether or not there is sufficient information contained on a PAR to determine if hospitalization had occurred for a case. By adding the two strata, the number of cases of serious injuries which could be available for sampling increased by 40%. In 1991, the two strata were added such that stratum J corresponds to late model year vehicles and stratum K corresponds to non-late model year vehicles.

In every one of the 171 police jurisdictions, a weekly sampling of PAR's is performed when a contractor by means of a computer program draws a sample of PAR's once all cases have been entered by a researcher. Afterwards, elements of the sample are transmitted to the researcher. This computer program is written in DELPHI for use in an Oracle database. Sampling worksheets upon which the DELPHI program is based were used when preparing this report in lieu of the DELPHI program to infer the sampling method of the PAR's.

The sizes of the samples depend on the availability of the researchers, consequently, the eventual sampling size is not always an optimum sampling size for achieving a prescribed precision in the estimates. Instead, the sizes of the samples of PAR's regardless of PSU will vary from week to week. In practice, the sampling of the PAR's is performed independently of the police jurisdiction since the operational sampling frame is formed by taking the union of the sampling frames of all sampled police jurisdictions within a PSU. It is from this combined sampling frame that the sample is drawn based on PPS sampling in which the measure of size of a PAR is the product of stratum weight by the size of the police jurisdiction with respect to its PJWGHT which is calculated by the method described in Section 2.2.1.

The stratum weights which are assigned to the PAR strata were created, in order to increase the number of cases of severe injuries in the sample. The weighting factors reflect the importance of these cases. The determination of the stratum weights was originally done in 1988 to produce a sample from which the proportion of late model year motor vehicle crashes to non-late model year motor vehicle crashes was 3 to 2. This ratio was changed to 7 to 3 when strata J and K were added.

The stratum weights which are called STRTWGHT are given in Table 5. In Table 5, PAR strata

Table 5: PAR Stratum Weight

PAR Stratum	A	B	C	D	E	F	G	H	J	K
STRTWGHT	400	400	175	25	7	3	2	1	400	300
Late Model Year		X		X		X		X		X
Non-late Model Year	X		X		X		X		X	

are designated Late Model Year and Non-late Model Year. Late model vehicles include the model year of the current year and the last three model years and the up-coming model year. Non-late Model vehicles include vehicles of a model year older than four years. [10]

2.4 Drawing a Sample

A discussion of the sampling of PAR's which are taken from the list of qualifying PAR's will begin with some definitions.

Definition 7. *1. Let $CONTDATE_j$ be the j^{th} date on which a researcher examines PAR_{ijkln} of PJ_{il} where n identifies the n^{th} PAR which is inspected on $CONTDATE_j$.*

2. Let $STRTWGHT_k$ be a stratum weight as given in Table 5 which is assigned to PAR stratum k.

3. Let $PARSTRT_{ijk}$ be the set of PAR's which belong to PAR stratum k of PSU_i and which were examined on $CONTDATE_j$. The subscript i identifies the PSU, k identifies the PAR stratum, and j identifies the CONTDATE.

4. Recall from Definition 2 that the subscript, l, identifies a PJ.

Upon being examined by a researcher, a PAR will not only be assigned to a PAR stratum according to the definitions appearing in Appendix III, but it is assigned by the Oracle database to a sequence number (SEQNUM) in sequential order as the PAR's are examined. Therefore, each qualifying PAR when it is listed in a sampling frame is assigned two identification numbers: the SEQNUM which is based on the order of discovery and a sampling frame identification number which is based on the time of the collision. This sampling frame identification number will be called the *sequence number*. The combination of these two identification numbers serves the purpose of ordering the PAR's within a stratum for PPS sampling. The process of sampling PAR's begins with their ordering. The PAR's are first sorted alphabetically by PAR strata and, then, within a PAR stratum by the sequence numbers in ascending order.

Definition 8. *1. Let $SEQUENCENUMBER_{ijn} = ACCTIMMM_{ijn} * 1000000 + ACCTIMHH_{ijn} * 10000 + SEQNUM_{ijn}$ be the sampling frame ordering number where ACCTIMMM and ACCTIMHH correspond to the minute and hour of the day on which the collision was reported to have occurred.*

For example, suppose a collision was reported to have occurred on 1 July at 1032, then ACCTIMHH=10 and ACCTIMMM=32. Suppose that the submitted PAR was the 38^{th} one which was examined by the researcher, then $SEQUENCENUMBER_{ijn} = 32100038$.

Once elements of the sampling frame are ordered by PAR stratum and then within a PAR stratum with respect to $SEQUENCENUMBER_{ijn}$, PAR's are selected using PPS sampling as illustrated in Example 2.

Example 2. *In Table 6, we see that the PAR's are ordered first according to PAR stratum and then according to the SEQUENCENUMBER within a given PAR stratum. For each PAR, a PARWGHT is*

*calculated by PARWGHT=STRTWGHT*PJWGHT. We see that the ordering of the PAR's is performed independently of the PJ's. The entries in Table 6 correspond to one PSU and to only one CONTDATE.*

Suppose that it has been decided that three PAR's are to be selected for investigation, then the sampling interval will be $\frac{98.90}{3} = 32.96667$. *To begin the process, we need a random number to determine the first element of the sample. Suppose .308 is that random number which a computer program generated, then the starting weight for sampling will be: .308*32.96667=10.15373. The first element to be selected for the sample, therefore, will be 32100038. The next element to be selected will be the one for which its cumsum is the smallest one which does exceed 10.15373+32.96667=43.1204. Thus 35170045 is the second element of the sample. Likewise, for the third and last element of the sample, 29070044 identifies the PAR with the smallest cumsum which exceeds* $10.15373 + 32.96667 + 32.96667 = 76.08707$.

Table 6: Sampling PAR's by PPS

SEQUENCENUMBER	SEQNUM	PJ	PJ Stratum	PAR Stratum	STRTWGHT	PJWGHT	PARWGHT	cumsum	select
32100038	38	4	4	E	7	1.98507	13.93	13.93	1
48090042	42	6	5	E	7	1.81538	12.74	26.67	0
1030004	4	9	6	F	3	2.11905	6.36	33.03	0
13140058	58	2	2	F	3	1.00000	3.00	36.03	0
35100026	26	1	1	F	3	1.00000	3.00	39.03	0
35170045	45	6	5	F	3	1.81538	5.46	44.49	1
57070059	59	2	2	F	3	1.00000	3.00	47.49	0
3000018	18	3	3	G	2	1.00000	2.00	49.49	0
16070012	12	9	6	G	2	2.11905	4.24	53.73	0
30020050	50	2	2	G	2	1.00000	2.00	55.73	0
45120021	21	1	1	G	2	1.00000	2.00	57.73	0
55140057	57	2	2	G	2	1.00000	2.00	59.73	0
4130046	46	6	6	H	1	1.81538	1.82	61.55	0
5000054	54	2	2	H	1	1.00000	1.00	62.55	0
5070030	30	10	7	H	1	3.31000	3.31	65.86	0
6080031	31	10	7	H	1	3.31000	3.31	69.17	0
14090019	19	3	3	H	1	1.00000	1.00	70.17	0
16160051	51	2	2	H	1	1.00000	1.00	71.17	0
20120022	22	1	1	H	1	1.00000	1.00	72.17	0
22170014	14	3	3	H	1	1.00000	1.00	73.17	0
25180052	52	2	2	H	1	1.00000	1.00	74.17	0
28120055	55	2	2	H	1	1.00000	1.00	75.17	0
29070044	44	6	5	H	1	1.81538	1.82	76.99	1
29140029	29	1	1	H	1	1.00000	1.00	77.99	0
30230056	56	2	2	H	1	1.00000	1.00	78.99	0
35100010	10	9	6	H	1	2.11905	2.12	81.11	0
35220033	33	10	7	H	1	3.31000	3.31	84.42	0
40110037	37	4	4	H	1	1.98507	1.99	86.41	0
41080053	53	2	2	H	1	1.00000	1.00	87.41	0
49230048	48	6	5	H	1	1.81538	1.82	89.23	0
50150007	7	9	6	H	1	2.11905	2.12	91.35	0
55180032	32	10	7	H	1	3.31000	3.31	94.66	0
57070006	6	9	6	H	1	2.11905	2.12	96.78	0
59040001	1	9	6	H	1	2.11905	2.12	98.90	0

2.4.1 PPS Sampling of PAR's

The method of PPS sampling is used to select police jurisdictions as well as to select PAR's. We will augment the definitions which we used for selecting PJ's by means of PPS sampling with a few more definitions.

Definition 9. *1. $SAMPLINGFRAME_{ij}$ is the set of all PAR's of the sampled PJ's which are examined on $CONTDATE_j$ and which describe a collision in which at least one automobile or one light truck has to be towed away due to damage and which had occurred on a highway in PSU_i.*

2. *Let $RAND_{ij}$ be the random number which is generated by a computer program for a $U(0,1)$ probability distribution for PSU_i. The same $RAND_{ij}$ applies to all elements of PSU_i on $CONTDATE_j$.*

3. *$PJWGHT_{iln}$ is the weight of PJ_{il} of PSU_i which is assigned to PAR_{ijklm}.*

4. *Let $PARWGHT_{ijkln} = STRTWGHT_k \ PJWGHT_{iln}$ for that PAR of PSU_i and PJ_{il} which was the n^{th} PAR which was inspected on $CONTDATE_j$.*

5. $\underset{n\leq\Lambda}{cumsum}(PARWGHT_{ijkln}) \ = \sum_{n=1}^{\Lambda} PARWGHT_{ijkln}$

6. *$CASELOAD_{ij}$ is the sampling size for the j^{th} week of PSU_i.*

7. $ORIGSAMP_{ij} = \underset{over\ all\ elements \in SAMPLINGFRAME_{ij}}{\sum} PARWGHT_{ijkln}$

8. $INTERVAL_{ij} = \frac{ORIGSAMP_{ij}}{CASELOAD_{ij}}$

9. $BEGIN_{ij} = INTERVAL_{ij} RAND_{ij}$

Once the sampling frame of the PSU has been ordered according to the SEQUENCENUMBER within each stratum, elements of the sample are drawn systematically beginning with $BEGIN_{ij}$ and at succeeding intervals, $BEGIN_{ij} + mINTERVAl_{ij}$, that is, the CDS sample for PSU_i and $CONTDATE_j$ is

Definition 10.

$$\mathcal{S}_{ij} = \{PAR_{ijkl\nu} | \nu = min\,\{\lambda | \ \underset{s\leq\lambda}{cumsum}(PARWGHT_{ijkls}) \ \geq BEGIN_{ij} + m\ INTERVAL_{ij}\}\ \forall m \leq CASELOAD_{ij}\}$$

2.5 Certainties

After the PAR's have been ordered alphabetically by strata and then within a stratum afterwards by the SEQUENCENUMBER, the process of sampling PAR's according to PPS sampling begins. Some PAR's will be classified as *certainties* when the probability of selection is 1. The process of identifying certainties is the following decision tree:

1. If any PAR listed in the ordered sampling frame has a PAR weight greater than or equal to the sampling interval, then it is selected with certainty.

2. Mark the case(s) as selected.

3. Recalculate the ORIGSAMP after removing the certainty case(s). Reduce the CASELOAD by the number of certainty cases selected.

4. Repeat Steps 1, 2, and 3 while using the revised ORIGSAMP and CASELOAD at each iteration until no more certainties have been discovered.

2.5.1 Sub-divided Police Jurisdictions

Some PJ's are sub-divided due to the extremely large number of cases to list. Within such a selected sub-division, half of the PAR's might be selected. For example, only odd numbered PAR's are selected. In other PJ's, every fifth PAR is selected. In these particular PJ's, the selection of sub-divisions constitutes another stage of sampling for a PAR. In spite of sub-dividing some PJ's, a total of 171 PJ's are eventually selected. The sub-sampling weights are listed in Table 7. These sub-weights are referenced in the program `CDSWGT.SAS` by the variable `ADJUST`.

Table 7: Criteria for Sub-dividing Police Jurisdictions

$ADJUST_{ijkl}$	Condition
2	if PSU=3
2	else if PSU=72 and $PJ \in \{1, 2, 3, 4, 5, 6\}$
5	else if PSU=72 and $PJ = 7$
2	else if PSU=79 and $PJ \in \{1, 7\}$
2	else if PSU=81 and $PJ \in \{1, 2\}$
1	otherwise

3 CDS Weights

The CDS weights are computed using the List Case file. This file contains the basic information about the cumulative collection of sampled PAR's for the year. A computer program produces from the List Case file the weights which are to be used for producing the national estimates of the information which is obtained from the researchers' investigations. The usual procedure for producing estimates from the CDS data is to calculate the frequency of each category of an investigation by means of RATWGT. The definition of RATWGT will be discussed later. Each observation contributes a value of one to the unweighted frequency counts, but when we use RATWGT, each observation is expanded to the national level so that it contributes the RATWGT weighted value for that observation.

As before, additional definitions will facilitate the description of the method of calculating CDS weights.

Definition 11. *1. Let $PARWGHT_{ijkln}$ be the sampling weight of PAR_{ijkln} which was drawn from the sampling frame with respect to PSU_i and $PARSTRT_{ijk}$ for the j^{th} CONTDATE.*

2. $SELECTED_{ijkln} = \begin{matrix} 1 & \textit{if } PAR_{ijkln} \textit{ was selected} \\ 0 & \textit{otherwise} \end{matrix}$

3. $CERTAINTY_{ijkln} = \begin{matrix} 1 & \textit{if } PAR_{ijkln} \textit{ was selected with probability 1} \\ 0 & \textit{otherwise} \end{matrix}$

4. $SELECT_{ij} = \sum_{over\ all\ elements \in SAMPLINGFRAME_{ij}} SELECTED_{ijkln}$

5. $CERT_{ij} = \sum_{over\ all\ elements \in SAMPLINGFRAME_{ij}} CERTAINTY_{ijkln}$

6. $CERTSUM_{ij} = \sum_{over\ all\ elements \in SAMPLINGFRAME_{ij}} PARWGHT_{ijkln} CERTAINTY_{ijkln}$

The variable $SELECT_{ij}$ is the same as the caseload for PSU_i on $CONTDATE_j$. $CERT_{ij}$ represents the number of cases of certainty which are contained in the sample for PSU_i on $CONTDATE_j$. $CERTSUM_{ij}$ is the total PARWGHT of all cases selected with probability 1 in PSU_i on $CONTDATE_j$.

We will define the step function, H(x), and three indicator variables as follows.

Definition 12. $H(x) = \begin{matrix} 1 & \textit{if } x > 0 \\ 0 & \textit{if } x \leq 0 \end{matrix}$

1. $CERTA_{ij} = H(SELECT_{ij} CERT_{ij})$

2. $NONC_{ij} = H(SELECT_{ij})(1 - H(CERT_{ij}))$

3. $OOPS_{ij} = 1 - H(SELECT_{ij})$

When $CERTA_{ij} = 1$, at least one PAR of PSU_i which is examined on $CONTDATE_j$ is selected with certainty while $NONC_{ij} = 1$ implies that no PAR's of PSU_i which are examined on $CONTDATE_j$ are selected with certainty. Supposedly, $CERTA_{ij}$ and $NONC_{ij}$ should account for all selected PAR's. Otherwise, it will be be flagged by $OOPS_{ij}$.

These many definitions are needed to define the following important quantity:

$$\begin{aligned} SI_{ijkln} \quad = \quad & PJWGHT_{iln} CERTAINTY_{ijkln} CERTA_{ij} + \\ & \frac{(ORIGSAMP_{ij} - CERTSUM_{ij})}{(SELECT_{ij} - CERT_{ij})} (1 - CERTAINTY_{ijkln}) CERTA_{ij} + \\ & \frac{ORIGSAMP_{ij}}{SELECT_{ij}} NONC_{ij} \quad SELECTED_{ijkln} \end{aligned} \tag{3}$$

Equation (3) reflects the three situations in which a PAR_{ijkln} can be found.

1. A PAR itself is selected with probability 1, i.e. with certainty.

2. The set of PAR's which have the same CONTDATE belonging to the same PSU has at least one of its members which was selected with certainty.

3. A PAR is not associated with any PAR which was selected with certainty.

If PAR_{ijkln} was selected with probability 1, then its SI_{ijkln} is the same as its $PJWGHT_{il}$. If the set of PAR's has members which were selected with certainty, they are removed from the other selected PAR's, so that the removal effectively reduces the sampling size by the number of certainties, that is, $SELECT_{ij} - CERT_{ij}$. Furthermore, the total sum of the PARWGHT's is reduced by the collective sum of the PARWGHT's of the certainties, that is, $ORIGSAMP_{ij} - CERTSUM_{ij}$. Finally, if a PAR is not associated with a certainty in any way, then $SI = \frac{ORIGSAMP_{ij}}{SELECT_{ij}} = \frac{sum\ of\ PARWGHT's}{Number\ of\ Selected\ PAR's}$.

The expansion factor to the national level is called the National Inflation Factor (NIF). Its definition is

Definition 13.

$$NIF_{ijkln} = PSUWGHT_i\ SI_{ijkln} \left(CERTAINTY_{ijkln} + \frac{(1 - CERTAINTY_{ij})}{STRTWGHT_k} \right) SELECTED_{ijkln} \quad (4)$$

In simplest terms, the NIF which is used in the operational CDS is basically the following:

$$NIF_{ijkln} = \frac{PSUWGHT_i}{CASELOAD_{ij}} \frac{\sum(STRTWGHT_k PJWGHT_{iln})}{STRTWGHT_k} \quad (5)$$

3.1 Benchmarking

The CDS weights which represent the product of basically three stages of sampling are rectified to known quantities in a method known as benchmarking. Recall that two PSU's are drawn from each of twelve PSU strata. The estimates which are published correspond to their respective PSU strata, called PSUSTRAT as shown in Table 2, and they correspond to groups of PAR strata called COLSTRAT as shown in Table 3.

The next step in calculating the CDS weights after the NIF_{ijkln}'s have been calculated is the implementation of a benchmarking method for rectifying the CDS weights to some known numbers of crashes in non-sampled and sampled PJ's with respect to PAR stratum and PSU. The process will produce the final weight denoted by RATWGT in the analytical files.

The final CDS weight is

Definition 14.

$$RATWGHT_{ijkln} = \frac{SUMCASEA_{m\zeta}}{SUMDENA_{m\zeta}} NIF_{ijkln} \quad (6)$$

where the quantities, $SUMCASEA_{m\zeta}$ and $SUMDENA_{m\zeta}$ depend on certain known quantities and the number of unselected PAR's from the current year's sampling frame. RATWGT is calculated with respect to PSUSRAT and to groups of PAR strata called COLSTRAT. In the computer program for calculating the final CDS weights, PSUSTRAT is denoted by the SAS variable, PSUGRP, as shown in Table 1. The definitions of PSUGRP and COLSTRAT appear in Tables 2 and 3.

Definition 15. *Let RN_{ik} be the number of PAR's of the union of non-sampled PJ's which were left over from the current year's sampling frame and which were assigned to PSU_i and to $PARSTRT_{ik}$. Values of RN are given in Appendix II for 2007.*

Let m designate the m^{th} PSUGRP as defined in Table 2, and let ζ designate the ζ^{th} COLSTRAT as defined in Table 3, then

$$SUMDENA_{m\zeta} = \sum_{k \in COLSTRAT_\zeta \; i \in PSUGRP_m} NIF_{i.k.} \tag{7}$$

$$SUMCASEA_{m\zeta} = \sum_{k \in COLSTRAT_\zeta \; i \in PSUGRP_m} (RN_{ik} + ADJUST_{i.k.})\, PSUWGHT_i \tag{8}$$

where $ADJUST_{i.k.} = \sum_{j=1}^{n_j} \sum_{l=1}^{n_l} ADJUST_{ijkl}$ and $ADJUST_{ijkl}$'s are defined by Table 7. ADJUST reflects the practice of sub-dividing a sampling frame of some PJ's by taking every other PAR from the sampling frame or every fifth PAR.

4 Trimming Weights

Though it is rarely employed, a process of trimming is performed. In order to mitigate excessively large $RATWGT_{ijkln}$'s, they are rounded not to exceed a ceiling which is determined by examining descriptive statistics of the RATWGT's. The excess amount which exceeds the ceiling is distributed uniformly over the remaining RATWGT's.

Bibliography

[1] *National Accident Sampling System, A Status Report, Volume I, Objectives of the National Accident Sampling System*, National Highway Traffic Safety Administration, Washington, D.C. 20590, 1978.

[2] *National Accident Sampling System, Pilot Study, Final Report, DOT HS-804 909*, National Highway Traffic Safety Administration, Washington, D.C. 20590, 1978.

[3] G. Binzer, H. J. Edmonds, , R. H. Hanson, D. R. Morganstein, and J. Waksberg, *NASS Estimation, Volumes 1: Technical Report, DOT HS 806 420*, Westat, Inc., 1650 Research Boulevard, Rockville, Maryland 20850, 1982.

[4] Nancy Bondy and Barbara Rhea, *Research Note, Reweighting of the Primary Sampling Units in the National Automotive Sampling System*, U.S. Department of Transportation, National Highway Traffic Safety Administration, 1200 New Jersey Avenue, S.E. Washington, D.C. 20590, 1997.

[5] National Safety Council, *Manual on the Classification of Motor Vehicle Traffic Accidents Sixth Edition (ANSI D 16.1 1996)*, National Safety Council, Itasca, Illinois, 1996.

[6] H. John Edmonds, Robert H. Hanson, David R. Morganstein, and Joseph Waksberg, *National Accident Sampling System Sample Design, Phases 2 and 3, Volumes I: Final Technical Report, DOT HS-805 274*, Westat, Inc., 1650 Research Boulevard, Rockville, Maryland 20850, 1979.

[7] _____, *National Accident Sampling System Sample Design, Phases 2 and 3, Volumes II: Exhibits, DOT HS-805 275*, Westat, Inc., 1650 Research Boulevard, Rockville, Maryland 20850, 1979.

[8] National Center for Statistics and Analysis, *NASS Brochure, http://www.nhtsa.dot.gov/portal/site/nhtsa/menuitem.331a23559ab04dd24ec86e10dba046a0/*, National Highway Traffic Safety Administration, Washington, D.C. 20590, 2009.

[9] R. H. Hanson, H. John Edmonds, L. Mohadjer, M. D. Rhoads, A. Chu, and D. R. Morganstein, *National Accident Sampling System Estimation, Final Report*, Westat, Inc., 1650 Research Boulevard, Rockville, Maryland 20850, 1985.

[10] Transportation Safety Institute, *NASS Sampling, Part One*, U.S. Department of Transportation, Oklahoma City, Oklahoma, 2000.

[11] Charles J. Kahane, *National Accident Sampling System, Selection of Primary Sampling Units, DOT HS-802 063*, National Highway Traffic Safety Administration, Washington, D.C. 20590, 1976.

[12] R. Kaplan and A. Wolfe, *Design for NASS: Supplemental Information for Planning the National Accident Sampling System, DOT-HS-801-989*, Highway Safety Research Institute, The University of Michigan, Ann Arbor, Michigan 48105, 1976.

[13] Eugene Lunn, Mike Brick, Ernst Meyer, Vern Roberts, Jim Hedlund, Jim Fell, Glenn Parsons, and Russell Smith, *National Accident Sampling System, A Status Report, Volume III, Implementation of NASS Subsystems*, National Highway Traffic Safety Administration, Washington, D.C. 20590, 1978.

[14] Ernst Meyer, *National Accident Sampling System, A Status Report, Volume II, Plan for a Pilot Study*, National Highway Traffic Safety Administration, Washington, D.C. 20590, 1978.

[15] J. O'Day, A. Wolfe, and R. Kaplan, *Design for NASS: A National Accident Sampling System. Volume I. Text.*, Highway Safety Research Institute, The University of Michigan, Ann Arbor, Michigan 48105, 1976.

[16] _____, *Design for NASS: A National Accident Sampling System. Volume II - Appendices, DOT-HS-4-00890*, Highway Safety Research Institute, The University of Michigan, Ann Arbor, Michigan 48105, 1976.

[17] Terry S. T. Shelton, *National Accident Sampling System, General Estimates System, Technical Note, 1988 to 1990, DOT-HS-807-796*, National Highway Traffic Safety Administration, Washington, D.C. 20590, 1991.

[18] Russell A. Smith, James Fell, and Charles J. Kahane, *FY 1977 Implementation of the National Accident Sampling System, DOT HS-802 260*, National Highway Traffic Safety Administration, Washington, D.C. 20590, 1977.

[19] Russell A. Smith and Eugene Lunn, *National Accident Sampling System, A Status Report, Volume IV, Implementation Schedule and Resource Requirements*, National Highway Traffic Safety Administration, Washington, D.C. 20590, 1978.

5 Appendix I: List of PSU's

Table 8:

PSU	PSUGRP	PSUSTRAT	Number of PJ Strata
2	9	3	8
3	1	1	9
4	9	3	8
5	5	2	10
6	1	1	9
8	5	2	13
9	7	8	10
11	10	6	7
12	6	5	7
13	10	6	7
41	3	7	4
43	11	9	6
45	7	8	3
48	11	9	8
49	3	7	2
72	2	4	7
73	6	5	7
74	2	4	7
75	8	11	5
76	12	12	8
78	12	12	5
79	4	10	4
81	8	11	8
82	4	10	2

6 Appendix II: Non-sampled PJ Counts

Table 9:

PSU	A	B	C	D	E	F	G	H	PSUGRP
2	1	6	5	30	46	162	182	417	9
3	0	1	121	205	527	809	871	1319	1
4	0	5	13	19	120	187	341	650	9
5	2	5	45	93	404	944	1139	2344	5
6	12	8	32	81	627	1668	552	1054	1
8	6	21	29	70	485	1059	1488	2580	5
9	3	7	78	141	155	379	384	921	7
11	0	0	1	0	0	7	13	33	10
12	0	9	8	16	64	162	189	434	6
13	0	0	0	1	1	6	5	22	10
43	0	1	1	4	22	82	90	198	11
73	1	4	4	44	95	274	311	850	6
75	0	1	3	2	10	24	37	103	8
76	0	0	1	3	6	5	15	15	12
78	3	3	2	7	6	19	20	65	12
79	1	2	37	43	1116	1208	2100	2461	4
81	1	1	9	21	65	237	397	1139	8

There are 17 PSU's in this table even though 24 PSU's were sampled between 2002 and 2007. Other PSU's did not contain any non-sampled PJ's; therefore there are no entries for them in this table. Those PSU's which do not contain any non-sampled PJ's are: 41, 45, 48, 49, 72, 74, and 82.

7 Appendix III: Definitions of PAR strata.

Stratum A - crashes in which at least one occupant of a towed CDS applicable late model year vehicle had a police reported injury of ”K” (fatal injury).

Stratum B - crashes not qualifying for Stratum A in which at least one occupant of a towed CDS applicable non-late model year vehicle had a police reported injury of ”K” (fatal injury).

Stratum J - crashes not qualifying for Strata A or B in which at least one occupant of a towed CDS applicable late model year vehicle had a police reported injury of ”A” (incapacitating injury) AND was transported to a treatment facility for treatment AND was admitted overnight to a hospital. If the crash involved more than one CDS applicable vehicle, at least two CDS applicable vehicles must be towed.

Stratum K - crashes not qualifying for Strata A, B or J in which at least one occupant of a towed CDS applicable non-late model year vehicle had a police reported injury of ”A” (incapacitating injury) AND was transported to a treatment facility for treatment AND was admitted overnight to the hospital. If the crash involved more than one CDS applicable vehicle, at least two CDS applicable vehicles must be towed.

Stratum C - crashes not qualifying for Strata A, B, J or K in which at least one occupant of a towed CDS applicable late model year vehicle had a police reported injury of ”A” (incapacitating injury) AND was transported to a treatment facility for treatment. If the crash involved more than one CDS applicable vehicle, then at least two CDS applicable vehicles must be towed.

Stratum D - crashes not qualifying for Strata A, B, J, K, or C in which at least one occupant of a towed CDS applicable non-late model year vehicle had a police reported injury of ”A” (incapacitating injury) AND was transported to a treatment facility for treatment. If the crash involved more than one CDS applicable vehicle, then at least two CDS applicable vehicles must be towed.

Stratum E - crashes not qualifying for Strata A, B, J, K, C or D in which at least one occupant of a towed CDS applicable late model year vehicle was transported to a treatment facility for treatment.

Stratum F - crashes not qualifying for Strata A, B, J, K, C, D or E in which at least one occupant of a towed CDS applicable non-late model year vehicle was transported to a treatment facility for treatment.

Stratum G - crashes not qualifying for Strata A, B, J, K, C, D, E or F which involve at least one CDS applicable late model year vehicle that was towed from the scene.

Stratum H - crashes not qualifying for Strata A, B, J, K, C, D, E, F or G which involve at least one CDS applicable non-late model year vehicle that was towed from the scene.

DOT HS 811 327
May 2010

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