U.S. Department
of Transportation

National Highway
Traffic Safety
Administration

People Saving People
http://www.nhtsa.dot.gov

DOT HS 809 541 **April 2003**

Technical Report

SAMPLING ISSUES IN REAR-END PRE-CRASH DATA COLLECTION

Published By:

National Center for Statistics and Analysis
Advanced Research and Analysis

This document is available to the public from the National Technical Information Service, Springfield, Virginia, VA 22161

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT HS 809 541		

4. Title and Subtitle	5. Report Date
Sampling Issues in Rear-End Pre-Crash Data Collection	April, 2003
	6. Performing Organization Code: NPO-121

7. Author(s)	8. Performing Organization Report No.
Santokh Singh, Ph. D.	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Rainbow Technology Inc. 17106 Thatcher Court Olney, MD 20832	11. Contract or Grant No.

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Mathematical Analysis Division, National Center for Statistics and Analysis National Highway Traffic Safety Administration, NPO-121 U.S. Department of Transportation 400 Seventh Street, S.W., Washington, D.C. 20590	NHTSA Technical Report
	14. Sponsoring Agency Code

15. Supplementary Notes

Discussions with Mr. Dennis Utter, Dr. Chou-Lin Chen (Mathematical Analysis Division, NHTSA), and Dr. Mike Goodman (Driver Vehicle and Simulation Division, NHTSA) were useful, as were the comments made by them. Proof reading by Mr. Tom Bragan is appreciated.

16. Abstract

A common type of crash that occurs on the roadways is the rear-end crash. Every year a large number of drivers are involved in such crashes. In order to develop effective crash countermeasures, it is important to have a better understanding of the driving behavior and performance of a driver prior to a rear-end crash. For that purpose, experiments need to be conducted in which the drivers can be observed in 'naturalistic' settings and data can be collected on the driver-related parameters.

This study discusses some of the sampling issues involved in the process of data collection in the above context. Contingency analysis is conducted to suggest criteria for stratifying the target population. A probabilistic approach is used for allocating the sample over the strata thus formed. An estimate of the number of vehicles needed to observe a specific number of rear-end crashes is obtained. This estimation problem is treated as the 'discrete waiting-time' problem. Additionally, Binomial probability distribution is used to estimate the number of drivers who would be involved in rear-end crashes as a result of deploying a certain number of vehicles. This estimate can be used to assess the potential of a given sample for acquiring the desired amount of information. When compared with some of the other methods of allocation (equal and proportional), the sample allocation criterion proposed in this study suggested much smaller sample size. Due to the random nature of rear-end crashes, the number of vehicles actually required for observing a certain number of rear-end crashes is likely to be large, while a smaller number may be deployable due to budgetary restrictions or other operational constraints. The sampling strategies are proposed for resolving the issues that may arise in such situation. The approach adopted in this study is fairly general and can be used to resolve the sampling issues in similar setups.

Two databases, the General Estimates System (GES) and the Fatality Analysis Reporting System (FARS), compiled by the National Highway Traffic Safety Agency (NHTSA), have been used in this study.

17. Key Words	18. Distribution Statement
crash involvement propensity, discrete waiting–time, equal allocation, inverse sampling, proportional allocation, rear-end, stratification, striking, struck	Document is available to the public through the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	7 No. of Pages	22. Price
Unclassified	Unclassified	27	

Table of Contents

EXECUTIVE SUMMARY

Background and objectives

The involvement of a driver in a rear-end crash and the manner in which his/her vehicle collides with other vehicle(s) depends not only on his/her perception of the complex scenario that emerges prior to the crash, but also on the pre-crash driving behavior, response to the imminent crash situation and performance in resolving the driving conflicts. Obviously, any effort directed towards crash countermeasures must start from data collection that can provide information about the driver-related parameters. This would further mean deploying vehicles on the roadways, which are equipped with certain recording devices, as well as making the voluntary drivers available. Due to the random nature of these crashes, the number of vehicles required to deploy for observing a certain number of them involved in such crashes may be large. On the other hand, due to budgetary restrictions and operational constraints, the sample size actually required may not be permissible. The effort should be, therefore, to make the best use of the available resources.

The objective of this study is to propose sampling strategies by which the maximum amount of information, both in terms of the number of rear-end crashes and coverage of the target population, could be obtained by deploying the minimum number of vehicles. Toward that end, sample stratification criteria are proposed. In addition, two estimates are proposed: (i) the number of vehicles that would be involved in rear-end crashes out of a certain number deployed, and (ii) the least number of vehicles that would be required for observing a specific number of vehicles involved in rear-end crashes. These estimates are required for implementation of the approach proposed in this study to resolve the sampling issues.

Data and methodology

Two databases, the General Estimates System (GES) of the National Automotive Sampling System (NASS) and the Fatality Analysis Reporting System (FARS), compiled by the National Highway Traffic Safety Administration (NHTSA) are used in the analysis.

The contingency analysis is used for proposing criteria that can be effectively used for stratifying the population of drivers for rear-end pre-crash data collection. Based on the likelihood ratio, a statistic is proposed for optimally allocating the sample size over the strata. The problem of estimating the number of vehicles/drivers required (sample size) until a given number of them is involved in rear-end crashes is treated as the ‘discrete waiting-time problem’. The sampling in this case is called ‘inverse sampling’. In addition, Binomial probability distribution is used for estimating the number of vehicles/drivers that would be involved in rear-end crashes out of a certain number engaged in data collection.

Conclusions and recommendations

The contingency analysis of GES and FARS data for the year 2000 showed that the driver attributes age and sex were associated with the driver’s involvement in a rear-end crash.

Accordingly, these two factors were considered as appropriate criteria for stratifying the population of drivers.

The statistic called Crash Involvement Propensity Index (CIPI) proved to be an efficient tool in optimally allocating the sample size over the strata, by making greater provision in the sample for the strata whose drivers were more prone to rear-end crash involvement. CIPI also provided a useful guideline for disbursing the target number over the strata, thus specifying an appropriate target number for each stratum.

CIPI-based sample allocation was compared for its efficacy with some of the other possible allocations: proportional and equal. The estimates obtained using different allocation methods showed that CIPI-based allocation was most efficient in both optimally disbursing the target number of drivers to be involved in rear-end crashes over the strata and allocating the sample when the sample size is fixed in advance.

Estimates of the numbers of vehicles that would be involved in rear-end crashes, as a result of deploying a certain number, showed that the sub-sample sizes corresponding to some of the strata would not generate sufficient amount of information in terms of the number of crash-involved drivers. This shortcoming was overcome by using the proposed sampling strategy. In this way, optimal sample could be designed that would not only be well representative of the drivers across the target population, but also be likely to produce sufficient information in terms of the drivers involvement in rear-end crashes. In yet another situation, the target numbers of rear-end crash-involved drivers from certain strata, as suggested by an allocation, may not be satisfactory. The sampling strategy proposed for handling such situation resulted in the smallest possible sample size.

The approach adopted in this study is general and can be used for resolving similar sampling issues involved in other data collection processes with similar setups.

1. Introduction

A common type of crash that occurs on the roadways is the rear-end crash, caused by one vehicle striking the rear of another vehicle when both vehicles are in the same traffic lane and are heading in the same direction. These crashes form a significant proportion of all crashes and involve a considerable number of drivers every year. Based on the databases, the General Estimates System (GES) of the National Automotive Sampling System (NASS) and the Fatality Analysis Reporting System (FARS), compiled by the National Highway Traffic Safety Administration (NHTSA), approximately 29.7% of all crashes were rear-end crashes in 2000. In terms of drivers crash involvement, of the 190,625,023 licensed drivers in 2000, reported by Federal Highway Administration (FHWA), approximately 2.2% were involved in rear-end crashes, making up 36% of all drivers involved in various types of crashes. These figures suggest the necessity of developing crash countermeasures that could prevent rear-end collisions. In this regard, it is becoming increasingly apparent that the development of any rear-end crash countermeasure would require a better understanding of the driving behavior and performance associated with the driver's response to driving conflict and imminent crash situations. This requires data collection in a "naturalistic" setting – crash situations as encountered by the drivers in their own cars, driving unobserved on the roadways. The vehicles deployed for this type of data collection must therefore be equipped with certain devices that could record the parameters related to the driving behavior and performance of a driver prior to a rear-end crash. In the subsequent discussion, these vehicles will be referred to as "experimental vehicles". The present study is focused on the sampling issues involved in rear-end pre-crash data collection. A probabilistic approach is used to formulate sampling strategies that can be used to resolve these issues. The sampling issues and specific objectives of the study are described in Section 2. Section 3 contains a brief introduction to the databases used in this study, as well as the rationale used in selecting the variables for statistical analysis. Some possible criteria for sample stratification and a statistic for sample allocation over the strata are proposed in Section 4. Section 5 is devoted to probabilistic formulation of the sampling issues. This section also provides analytical details of the estimates required for the sampling strategies proposed in this study. The implementation of these strategies is demonstrated through examples in Section 6, followed by discussion of the results and recommendations in Section 7. The contingency tables and the analytical details of the statistic Crash Involvement Propensity Index (CIPI) [2] are included in the Appendices (Section 8 and Section 9, respectively). The references used in this study are listed in Section 10.

2. Sampling issues and objectives of the study

In real-life driving conditions, the parameters related to the driving behavior and performance of a driver are determined by the complex scenario that emerges prior to a rear-end crash. While data collection is crucial for acquiring information on the driver-related parameters, an efficient sample design is important from the point of view of conserving resources that are required in terms of the experimental vehicles, as well as the voluntary drivers that need to be made available to drive these vehicles. With the above objective in mind, two aspects of the sample design need to be considered. If the aim is to collect data on a specific number of rear-

end crashes, then the vehicles comprising the sample have to be kept deployed until the required number of these have been involved in rear-end crashes. In statistical terms, this sampling procedure is called "inverse sampling". Since the emergence of scenarios resulting in the occurrence of rear-end crashes is random, the inverse sampling may require a large number of vehicles in order to observe a specific number of them involved in the rear-end crashes. On the other hand, due to budgetary restrictions on the number of experimental vehicles or other operational constraints, such as the availability of voluntary drivers, only a fixed number of vehicles may be available for the experimentation (data collection). In other words, the sample size is fixed in advance. One must then look at the second aspect of the sample design. In this case, it is important to estimate the number of rear-end crashes that would occur if the pre-specified number of experimental vehicles were deployed for data collection. This estimate can help in assessing the amount of information that a sample of given size would generate. If the given sample size is not large enough to produce a sufficient amount of data on the rear-end crashes, some sampling strategy needs to be used by which one could make the best use of the available number of experimental resources.

The objective of the present study is to propose sampling strategies that are optimal both in terms of the number of rear-end crashes and the content of information about the driver-related parameters across the population of drivers. In this context, it is important to keep in mind that an experimental unit in the current data collection process is comprised of a vehicle and a voluntary driver. A diligent selection of drivers from the target population is crucial for arriving at an optimal sample design. One of the ways to do this is to stratify the population using an appropriate criterion and select drivers from the strata proportional to the crash involvement propensity of drivers in each stratum.

In the subsequent sections, for the purpose of rear-end pre-crash data collection, we will

i. propose efficient criteria for stratification of the target population of drivers,

ii. propose a criterion that can be used to optimally allocate the sample over the strata,

iii. estimate the sample size that is large enough to be able to observe a specific number of drivers who would be involved in the rear-end crashes,

iv. estimate the number of drivers who are likely to be involved in the rear-end crashes as a result of a specific number engaged in data collection.

3. Data sources and target variables

The statistical analysis conducted and the resulting conclusions made in this study are based on the information/data retrieved from the following sources:

1. Age/Sex distribution of licensed drivers for 2000, reported by the FHWA

2. Drivers involved in rear-end and other crashes in 2000, reported in GES

3. Drivers involved in fatal crashes in 2000, reported in FARS

While GES obtains its data from a nationally representative probability sample selected from the estimated police reported crashes, FARS contains the data only from the files that document all qualifying fatal crashes. For that reason, cases with fatal crashes were used from FARS data in lieu of the fatal crashes estimated in GES data.

Keeping in mind that our interest is in the driving behavior and performance of the driver prior to a rear-end crash, the factor (variable) that most deserves attention is the *Manner of collision* in a crash. Of the several manners of collision coded in these databases, we will focus on *Rear-end* collision. Since the manner of collision of a vehicle amounts to the involvement of its driver in a certain type of crash, and our interest is in the rear-end crashes, for the subsequent analysis, we define a new variable *Crash event* as

$$Crash\ event = \begin{cases} \text{Rear - end, if the manner of collision is rear - end} \\ \\ \text{Other , if the manner of collision is other than the rear - end} \end{cases}$$

Last but not least, the perception of the circumstances surrounding a crash as well as the driving behavior and performance of a driver prior to a crash seem to be related to driver attributes age and sex. Accordingly, our focus in this study will be on two more variables *Age* and *Sex*.

4. Stratification and sample allocation criteria

The basic aim in any data collection process is to acquire a maximum amount of desired information at the minimum cost and effort. Therefore, whether or not there is a restriction on the sample size that can be used in a given situation, it can help a great deal in achieving this aim if the target population is first stratified using an appropriate criterion and then an efficient criterion is used to optimally allocate the sample over the strata thus formed.

4.1. Stratification criteria

While the involvement of a driver in a crash depends on his/her perception of the complex scenario that emerges prior to a crash, a driver's pre-crash driving behavior and performance plays an important role in resolving the driving conflicts. This suggests that the driver attributes age and sex may be two of several factors contributing to the rear-end crash involvement of a driver. These attributes can, therefore, be considered as possible factors for stratification of the target population. Nevertheless, using these factors for this purpose will make sense only if there is an evidence of the association between *Age/Sex* and *Crash event*.

Contingency analysis was performed for testing the association between driver's *Age* and *Crash event*. This is one of the useful techniques that can be used for analyzing the data that can be meaningfully classified in a contingency table, such as Table A.1 (Appendix A). In order to test the independence between driver attribute age and crash event, the drivers were classified using the following mode.

- Classification of drivers based on age:

  A_1: Age group 1 (younger than 18)
  A_2: Age group 2 (18 to 24)
  A_3: Age group 3 (25 to 44)
  A_4: Age group 4 (45 to 64)
  A_5: Age group 5 (older than 64)

With this criterion of classification in place, the contingency analysis of GES data for the year 2000 was carried out to test the hypothesis: *there is no association between driver's Age and Crash event*. Since the collection of GES data is based on three-stage sampling, the statistical software SUDAAN 8.01 was used for this purpose, which takes into account the underlying sampling design of the data being used in the analysis. The results are presented in the contingency table (Table A.1, Appendix A.) that yield $\chi^2 = 159.2$ with 4 degrees of freedom. The 95th percentile 9.49 of χ^2 distribution (with 4 degrees of freedom) being far less than 159.2, the hypothesis of no association is discredited, thereby indicating that there is a strong evidence of an association between driver's *Age* and *Crash event*.

Based on GES and FARS data for the year 2000, it can be seen that among drivers older than 17, 18 to 24 year old drivers (considered as young drivers in this study) have the highest rate of involvement in rear-end crashes (1,964 per 100,000 young drivers). The drivers from this age group can, therefore, provide more data on the driver-related parameters as compared to the drivers belonging to other age groups. However, before a sample is designed from this age group, it is worth investigating if the attribute sex should be used for stratifying the population of young drivers. Accordingly, the young drivers were classified using the following mode.

- Classification of young drivers based on *Sex*:

  Y_m: Young male (18 to 24 male drivers)
  Y_f: Young female (18 to 24 female drivers)

The contingency analysis was carried out to test the hypothesis: *there is no association between Sex of the young driver and Crash event.* The results are presented in the contingency table, Table A.2, (Appendix A) that yield $\chi^2 = 6.96$ with 1 degree of freedom. Since the 95th percentile 3.84 of χ^2 distribution with 1 degree of freedom is less than 6.96, we would reject the hypothesis of independence between *Sex* of the young driver and *Crash event* and conclude that the young driver's sex has some bearing on his/her involvement in a rear-end crash.

The statistical evidence of the association between *Crash event*, on the one hand, and driver's *Age* or young driver's *Sex* on the other (as shown through the above analysis) gives a strong reason to use these driver attributes as the stratification criteria in designing a sample for rear-end pre-crash data collection.

In the subsequent discussion, the classes of drivers defined on the criteria of *Age* and *Sex* will be referred to as the attribute-based classes.

4.2. Sample allocation criterion

The stratification based on age/sex can be effectively used in the present context, if the sample is designed in such a way that more drivers are included from the strata that consist of drivers who are more prone to rear-end crash involvement. Once this is done, the resulting sample would not only increase the likelihood of more drivers involved in rear-end crashes and hence yield more data on the driver related parameters, but also provide the desired information across the target population.

Generally speaking, if the population of drivers is stratified over *M* strata on a certain criterion, what one needs to look for is the likelihood (crash involvement propensity) of a driver belonging to the *i*th stratum being involved in a rear-end crash relative to that of the drivers from other *M-1* strata. In order to arrive at a suitable measure of the crash involvement propensity of drivers belonging to a stratum as compared to other strata, it is important to consider the occurrence of rear-end crash-involved drivers in a stratum relative to the occurrence of its drivers in the entire population of drivers. The important information that one needs in this context is an answer to the question: Given that a driver selected at random is from a certain stratum, what is the probability that he/she would be involved in a rear-end crash? These probabilities can then be combined into the statistic ϕ_i, called *Crash Involvement Propensity Index* (CIPI) [2], given by (for analytical details refer to Appendix B)

$$\phi_i = \frac{\frac{C_i}{S_i^2}}{\sum_{j=1}^{M}\left(\frac{C_j}{S_j^2}\right)}, \quad i = 1, 2, \ldots, M, \tag{1}$$

where

C_i is the number of rear-end crash-involved drivers belonging to the *i*th stratum (*i*th subpopulation),

S_i is the number of drivers in the *i*th stratum, i.e., the size of the *i*th subpopulation; $S_i > 0$, $S_1 + S_2 + \ldots + S_M = N_T$ (size of the population of all drivers),

M is the number of disjoint strata that are exhaustive of the population of drivers.

Note that the numerator in (1) takes into account the likelihood or conditional probability (conditional on stratum) of a driver belonging to the *i*th stratum being involved in a rear-end crash (Appendix B), while the denominator is the normalizing quantity. Obviously, the index ϕ_i satisfies the inequality $0 \le \phi_i \le 1$. The statistic CIPI given in equation (1) provides a measure of the propensity of drivers belonging to a certain stratum of being involved in the

rear-end crashes, relative to that of the drivers of other strata.

Used as the constant of proportionality for allocating the sample over the strata, this index can split the sample size n^* in such a way that larger sub-sample sizes are assigned to the strata that consist of drivers with higher rear-end crash involvement propensity. Specifically, the strata sample sizes can be computed from the relation

$$n_i = n^* \left\{ \frac{\frac{C_i}{S_i^2}}{\sum_{j=1}^{M} \left(\frac{C_j}{S_j^2} \right)} \right\}, \quad i = 1, 2, \ldots, M \tag{2}$$

where n_i is the sample size for the ith stratum and n^* is the total sample size. Obviously, $\sum_{i=1}^{M} n_i = n^*$.

5. Probabilistic look at the sampling issues

As mentioned earlier, there are mainly two sampling issues involved in designing a sample for the rear-end pre-crash data collection: (i) to estimate the sample size required for observing a target number of drivers involved in rear-end crashes, and (ii) to utilize a sample of pre-specified size for obtaining the maximum possible information in terms of the rear-end crash involvement. Before methods can be developed to resolve these issues, it is important to remember that a driver's involvement in a rear-end crash is one of the several road events that are random. This allows us to formulate the current sampling problem, probabilistically.

For that purpose, define the events E_1 and E_2 as

E_1: Driver is involved in a rear-end crash,

E_2: Driver is involved in a crash other than the rear-end crash or is not involved in any type of crash.

The definitions of these events suggest that each driver in an attribute-based class can be categorized in one of the two categories C_1 and C_2, depending on the occurrence of the events E_1 and E_2, respectively. For instance, if a driver is involved in a rear-end crash, then he/she will be considered as belonging to C_1. Similarly, if a driver is involved in a head-on collision or is not involved in any type of crash, then he/she will be considered as belonging to C_2. This categorization of drivers will be referred to as the event-based categorization and is used for categorizing each age- and sex-based class of drivers defined in Section 4. The resulting categorization is shown in Figure 1, where Figure 1(a) shows the event-based categorization

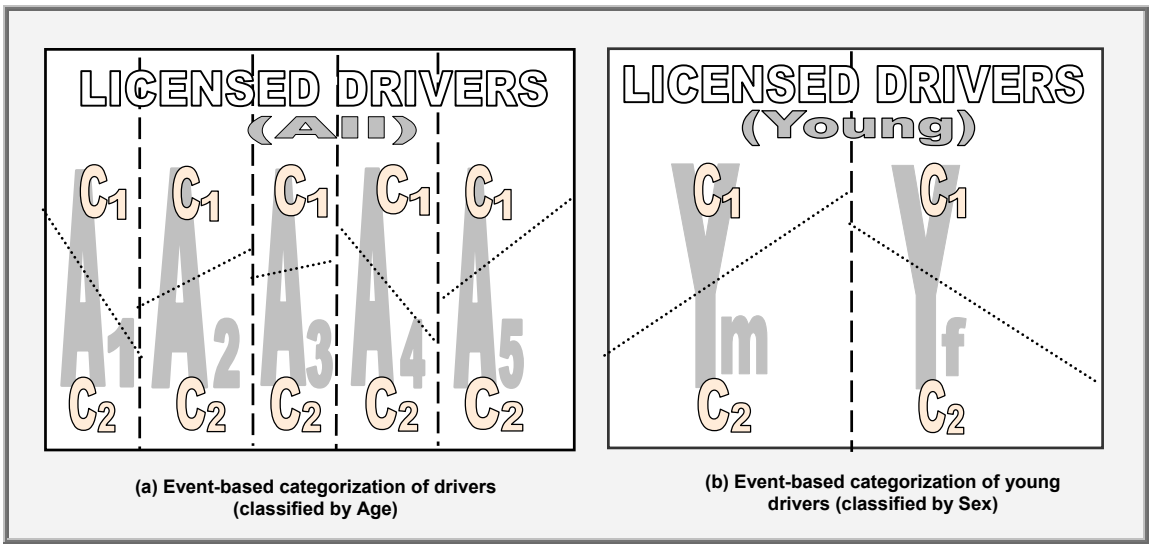

Figure1. Categorization of drivers (classified by *Age* and *Sex*), based on *Crash event*: Rear end crash involvement or otherwise.

of drivers classified into age groups A_1, A_2, A_3, A_4, and A_5, and Figure 1(b) shows the event-based categorization of young male (Y_m) and young female (Y_f) drivers. Note that the areas demarked in Figure 1 are merely representative of the class/category of drivers and not of their actual sizes. In the subsequent analysis, the age- and sex-based classes A_1, A_2, A_3, A_4, A_5, Y_m, Y_f, will be referred to also as subpopulations.

5.1. Inverse sampling

Consider, first, the situation where the objective is to determine the number of vehicles that need to be included in the fleet of experimental vehicles so that at least k of the drivers would be involved in rear-end crashes. Keeping in view the fact that the drivers in each subpopulation are further subdivided into the event-based categories C_1 and C_2, the sampling process can be thought of as drawing objects (drivers) from a box (class) containing N objects of two types (C_1 and C_2) until the target number (k) of objects of one type (drivers from C_1) are included in the sample.

Due to the uncertainty inherent in the road events, each of the event-based categories of a subpopulation can be associated with a certain probability of occurrence in the object drawing process, described above. Specifically, let

p be the probability that a driver is involved in a rear-end crash, i.e., the probability that the driver belongs to C_1,

q be the probability that a driver is involved in a crash other than the rear-end crash or is not involved in any type of crash, i.e., the probability that the driver belongs to C_2 ($q=1-p$).

Equivalent to the box problem, as described above, the data collection in the present case has to continue until *k* drivers are involved in rear-end crashes; the probability being p that a driver would be involved in a rear-end crash. This has repercussions in that the termination of data collection depends on the number (*k*) of drivers involved in rear-end crashes that have occurred up to and including a crash and not on any other road event. These facts provide sufficient reason to treat the current sample size problem as a *discrete waiting-time* problem.

For that purpose, we define the event *E* as

E = {exactly *k* drivers are involved in rear-end crashes, i.e., belong to C_1}.

Obviously, the event *E* can happen only when a crash occurs that ends with exactly *k* drivers involved in rear-end crashes. A typical sequence of the drivers involved in all types of road events, before 3 of them, for example, are involved in rear-end crashes, would look like the one shown in Figure 2. Recalling the definition of the events E_1 and E_2, it is easy to see that

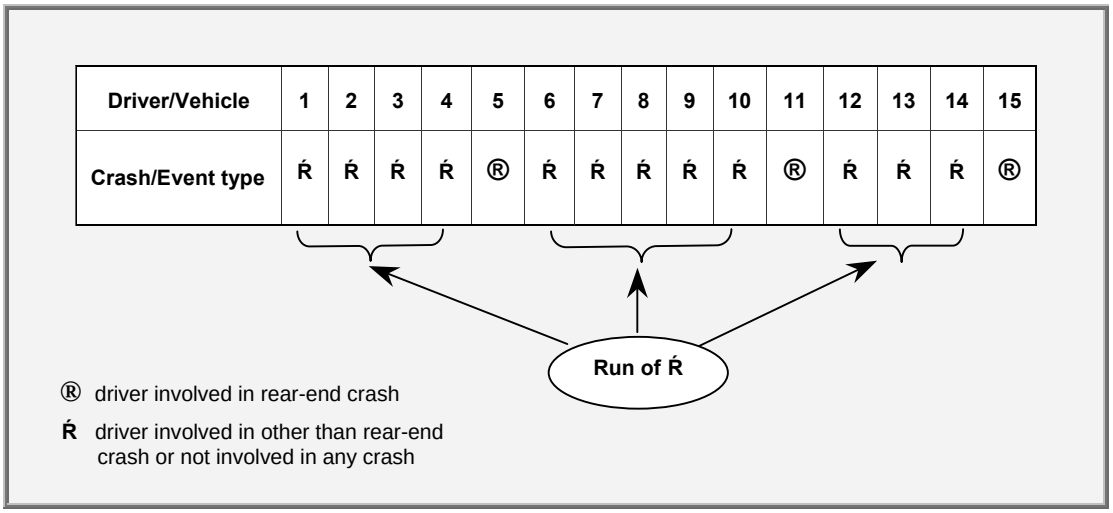

Figure 2. Sequence of crash events required to wait through for the involvement of $k = 3$ drivers in rear-end crashes.

while the occurrence of the event *E* is governed by the number of occurrences of E_1, the occurrence of E_2 decides the length of the sequence of road events that would be required to have happened before the event *E* happens. It follows that the number of drivers (vehicles) required for the occurrence of *E* is a random variable that assumes values: $k, k+1, k+2, \ldots, Nq;\ 1 \le k \le Np$, where N is the total number of drivers in a population/subpopulation.

If X is the number of drivers one has to wait through before *k* of them are involved in rear-end crashes, then the probability distribution of X can be described by the *Hypergeometric waiting-time distribution*, defined as [1]

$$Prob(X = x) = \frac{\binom{x-1}{k-1}\binom{N-x}{Np-k}}{\binom{N}{Np}},\ x = k, k+1, \ldots, Nq;\ 1 \le k \le Np. \qquad (3)$$

Using the probability distribution from equation (3), the expected number of drivers required in order to have at least *k* of them from C_1 is given by

$$\hat{n}(N, k, p) = \sum_{x=k}^{Np} x \frac{\binom{x-1}{k-1}\binom{N-x}{Np-k}}{\binom{N}{Np}}$$

which simplifies to

$$\hat{n}(N, k, p) = \frac{k(N+1)}{Np+1}. \qquad (4)$$

and the variance of the sample size is given by

$$V(k, p) = \frac{Nq(Nq-1)k(k+1)}{(Np+2)(Np+1)} + \frac{k(2k+1)(N+1)}{Np+1} - \frac{k^2(N+1)^2}{(Np+1)^2} - k(k+1) \qquad (5)$$

that can be used to calculate the confidence interval for the number of drivers required in order to observe at leas *k* of them involved in rear-end crashes.

5.2. Binomial sampling

Consider now the situation when the sample size n^* is fixed in advance and the requirement is to find the expected number $\hat{k}$ of drivers who would be involved in rear-end crashes, i.e., would be in C_1. Obviously, in this case, during the experimentation, each of the n^* drivers would fall either in C_1 or in C_2 with *p* as the probability of a driver being in C_1. This phenomenon can, therefore, be described by Binomial probability distribution with parameters n^* and *p*. Accordingly, the expected number of drivers who would be involved in rear-end crashes in a sample of size n^* is given by [1]

$$\hat{k}(n^*, p) = n^* p. \qquad (6)$$

with the variance

$$V(n^*, p) = n^* pq \qquad (7)$$

that can be used to calculate the confidence interval for the number of drivers involved in rear-end crashes out of a sample of size n^*.

6. Sample designs

Whether or not the sample size is fixed in advance, our main concern in pre-crash data collection is to obtain maximum information on the driver-related parameters in rear-end crashes at the minimum cost. Based on the previous analysis, the most important information that can be utilized to achieve this objective is the differential that exists in terms of the crash involvement propensity of drivers belonging to different age- or sex-based classes of drivers. The attribute-based classes will henceforth be referred to as strata. The most efficient way of utilizing this differential is to allocate more drivers (out of the pre-specified sample size) to those strata that have a higher crash involvement propensity. Similarly, more drivers are expected (out of the target number required to be involved in rear-end crashes) from those strata that have higher crash involvement propensity. In either case, some of the strata may not generate as much information, in terms of the number of drivers involved in rear-end crashes, as is required in order to arrive at useful conclusions. This may happen because some of the strata may ask for larger numbers of vehicles to be deployed than the corresponding numbers that can actually be allocated due to the over all restriction on the sample size. However, an efficient sampling strategy can help a lot in making up the discrepancy. In the subsequent sections, the sampling strategies are proposed for this purpose and are demonstrated through examples. Two situations are considered: (i) when the number of drivers that must be involved in rear-end crashes, before the data collection is stopped, is pre-specified and the aim is to obtain sufficient information on crash-involved drivers from each attribute-based stratum, and (ii) when the sample size is pre-specified and the aim is to allocate drivers over the attribute-based strata so that each sub-sample can produce sufficient information in terms of the rear-end crash involvement of drivers. While situation (i) can be handled by inverse sampling, both binomial and inverse sampling are required for resolving the sampling issues arising in situation (ii). The following two examples not only demonstrate the implementation of the proposed sampling strategies, but also compare CIPI-based allocation with some of the other possible allocations: equal and proportional.

6.1. Case 1. Sample designs with the target number, *k*, of drivers to be involved in the rear-end crashes

Consider the situation where the data on pre-crash driver-related parameters need to be collected on *k* drivers involved in rear-end crashes and there is no restriction on the number of vehicles/drivers needed for achieving this objective. The proposed approach to arrive at an optimal sample design in this situation is demonstrated through examples in the following sections.

GES and FARS data for the year 2000 were used to estimate the sample size required for observing *k* (=10) drivers involved in rear-end crashes. Due to the anticipated operational difficulties, Age group 1 and Age group 5 were excluded. Specifically, two target populations were considered: 18 to 64 year old drivers (Age group 2, Age group 3, and Age group 4), and 18 to 24 year old (young) drivers. As suggested by the contingency analysis in Section 4.1, two stratification criteria were used: age of the driver in the first population and driver's sex in the second. The target number *k* of drivers required to be involved in rear-end crashes was

partitioned into M numbers, $k_1, k_2,\ldots,k_M$ ($\sum_{i=1}^{M} k_i = k$), with k_i representing the target number of rear-end crash-involved drivers from stratum i. This was done in four ways: (i) k_i's determined by CIPI, (ii) equal k_i's, (iii) k_i's proportional to strata sizes, and (iv) k_i's adjusted using the sampling strategy. The steps required to arrive at an optimum sample size through the sampling strategy are described in Figure 3. The strata sample sizes and hence the total sample size estimated by CIPI and the sampling strategy use inferences made in the previous sections as well as the estimates obtained there.

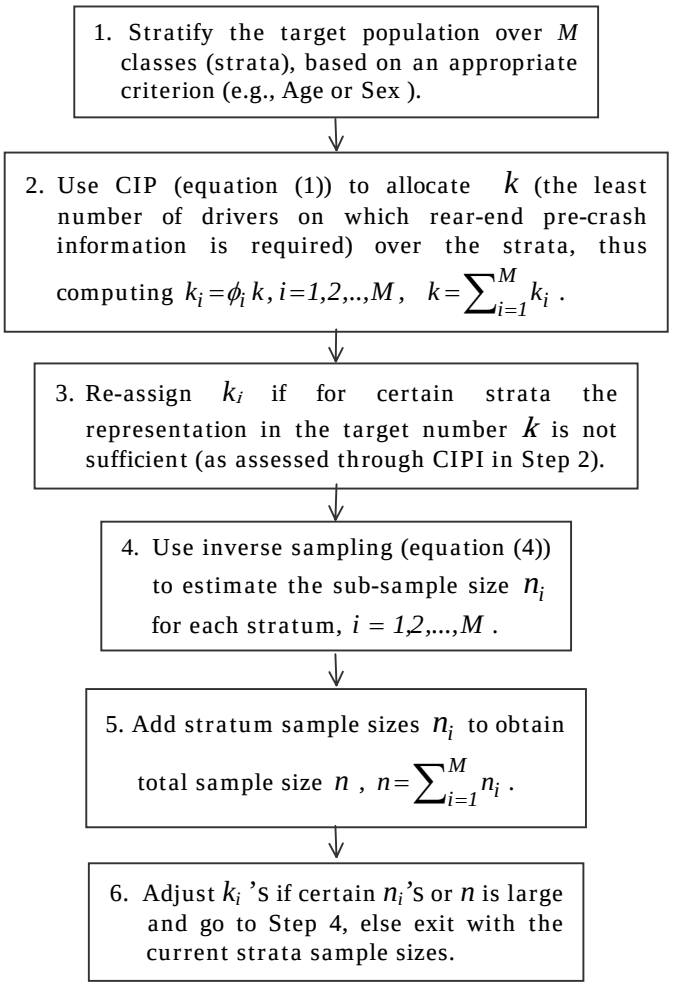

Figure 3. Sampling strategy when the target number, *k*, of drivers involved in rear-end crashes is to be observed.

The results, based on one-year crash statistics, are presented in Table 1 (for age-based stratification, M=3) and Table 2 (for sex-based stratification, M=2). The last three columns (columns 5, 6, and 7) in these tables, respectively, represent the number of drivers likely to be involved in rear-end crashes from each stratum (k_i), the required stratum sample size (n_i) and its 95% confidence interval.

6.1.1. Example 1: Sample design with the target number, *k*=10, in sampling from a population stratified by driver's age

In this example, the stratification of 18 to 64 year old drivers was done on the criterion of age. The estimated sample sizes and the corresponding 95% confidence intervals are presented in Table 1. These results show that, if CIPI is used as the sample allocation

Table 1. Sample designs for observing the target number, *k=10,* of drivers involved in rear-end crashes, using CIPI, equal and proportional allocations and the sampling strategy in sampling from a stratified (by age) population of drivers

Allocation criterion	Stratum of drivers (i)	Probability rear-end crash involvement (p_i)	Constant of proportionality (α_i)	Number of drivers involved in rear-end crashes ($k_i = \alpha_i.k$)	Stratum sample size (n_i)	95% Confidence interval [$n_{i\,LOWER}$, n_{iUPPER}]
CIPI	1. Age group 2	0.03742	0.74949	8	213	[68, 359]
	2. Age group 3	0.02393	0.13294	1	42	[2, 123]
	3. Age group 4	0.01575	0.11758	1	64	[1, 187]
	Total			10	319	
EQUAL	1. Age group 2	0.03742	0.33333	3 + 1	107	[4, 210]
	2. Age group 3	0.02393	0.33333	3	125	[3, 266]
	3. Age group 4	0.01575	0.33333	3	191	[3, 404]
	Total			10	423	
PROPORTIONAL	1. Age group 2	0.03742	0.13721	1	37	[0, 97]
	2. Age group 3	0.02393	0.49472	5	207	[3, 387]
	3. Age group 4	0.01575	0.36807	4	233	[3,471]
	Total			10	477	
SAMPLING STRATEGY	1. Age group 2	0.03742	*n*	5	134	[19, 249]
	2. Age group 3	0.02393	*n*	3	125	[2, 265]
	3. Age group 4	0.01575	*n*	2	127	[1,302]
	Total			10	386	

Source: National Center for Statistics and Analysis, NHTSA, GES and FARS 2000, FHWA

n not required

criterion, then, of the 10 drivers required to be involved in rear-end crashes, 8 would be from Age group 2, requiring 213 drivers, and 1 each from Age group 3 and Age group 4, requiring, respectively, 42 and 64 drivers to participate. Thus, with age as the stratification criterion and CIPI as the allocation criterion, the total number of vehicles/drivers required for data collection adds up to 319.

For the purpose of comparison, two other methods of allocation were also considered: equal and proportional. The results presented in Table 1 show that with the target numbers $k_1 = 4$, k_2

= 3, and k_3 = 3 for the three strata, larger number of drivers, 125 and 191, would be required, respectively, from Age group 3 and Age group 4. This is obviously due to the fact that Age group 3 and Age group 4 have lower crash involvement propensity as compared to Age group 2 and yet the samples selected from them are supposed to produce almost the same number involved in the rear-end crashes. This in turn raises the requirement of total number of vehicles/drivers in the sample to a larger number (423) as compared to CIPI-based allocation that requires 319 drivers.

The target number (k = 10) of crash-involved drivers was also allocated using proportional allocation (i.e., proportional to strata sizes). The results presented in Table 1 show that with this allocation, only 1 driver is expected to be involved in a rear-end crash out of 37 selected from Age group 2. Age group 3 and Age group 4 need to contribute, respectively, 207 and 233 drivers to the sample in order to observe, respectively, 5 and 4 drivers involved in rear-end crashes. Thus, with proportional allocation, 477 vehicles/drivers would be required in order to observe 10 drivers involved in rear-end crashes. This number is larger than the one suggested by equal allocation (423) and much larger as compared to 319 suggested by CIPI-based allocation.

In certain situations, based on a sample allocation, the contributions of some strata to the target number k may not be satisfactory and k_i's need to be adjusted. In such situations, it is advisable to use a sampling strategy rather than determining k_i's arbitrarily. Assume for the time being, that in the present case, this is the situation. Using the sampling strategy proposed in Figure 3, optimal values of k_1, k_2, and k_3 were determined as 5, 3, and 2, respectively. The estimates of the corresponding sample sizes required from Age group 2, Age group 3 and Age group 4 were obtained as 134, 125, and 127, respectively. After adjustment of k_i's, the total sample size requirement for observing 10 drivers involved in rear-end crashes is much larger (386) as compared to CIPI-based (319). This is obviously due to the requirement of higher rear-end crash involvement of drivers from the strata which, otherwise, have low crash involvement propensity.

The above results show that of the three sample allocations used in this example, the one based on CIPI requires the smallest number (319) of vehicles/drivers to be engaged in data collection. The proposed sampling strategy increases the likely contributions of Age group 3 and Age group 4 from 1 each (as suggested by CIPI) to, respectively, 3 and 2 to the subpopulation of rear-end crash-involved drivers. This, however, increases the total sample size requirement from 319 (CIPI-based allocation) to 386 which is still less than the sample size based on equal (423) and proportional (477) allocations.

6.1.2. Example 2: Sample design with the target number, *k*=10, in sampling from a population stratified by driver's sex

It can be seen in Table 1 (previous section) that among the three age groups, Age group 2 has the highest crash involvement propensity, 0.74949, used as constant of proportionality in CIPI-based allocation. This suggests that if the voluntary drivers are selected from only Age group 2, it may be economical in terms of the number of vehicles/drivers required for data collection. Accordingly, the sub-population of young drivers was considered as the target

population and stratified by driver's sex. In order to design sample, the target number *k* (=10) of young drivers required to be involved in rear-end crashes was split over the two strata (male and female drivers), using three methods of allocation: CIPI, equal, and proportional. The results are presented in Table 2.

Table 2. Sample designs for observing the target number, *k=10,* of drivers involved in rear-end crashes, using CIPI, equal and proportional allocations in sampling from a stratified (by sex) population of young drivers

Allocation criterion	Stratum of drivers (i)	Probability rear-end crash involvement (p_i)	Constant of proportionality (α_i)	Number of drivers involved in rear-end crashes ($k_i = \alpha_i.k$)	Stratum sample size (n_i)	95% Confidence interval [$n_{i\,LOWER}$, n_{iUPPER}]
CIPI	1. Young male	0.04199	0.55045	6	143	[31, 255]
	2. Young female	0.03260	0.44955	4	123	[4, 241]
	Total			10	266	
EQUAL	1. Young male	0.04199	0.50000	5	119	[17,221]
	2. Young female	0.03260	0.50000	5	153	[21, 286]
	Total			10	272	
PROPORTIONAL	1. Young male	0.04199	0.51268	5	119	[17,221]
	2. Young female	0.03260	0.48732	5	153	[21, 286]
	Total			10	272	

Source: National Center for Statistics and Analysis, NHTSA, GES and FARS 2000, FHWA

The CIPI-based allocation suggests that in order to observe 10 young drivers (6 male and 4 female) involved in rear-end crashes, the sample needs to be comprised of 143 young male and 123 young female drivers. On the other hand, if equal number (5) of young male and young female drivers involved in rear-end crashes are to be observed, then 119 drivers will have to be selected from young male drivers and much larger (153) from young female drivers. The proportional allocation results in the same estimates of the strata sample sizes as the equal allocation. The results show that the allocation based on CIPI requires the smallest sample size for observing the target number (10) of drivers involved in rear-end crashes.

6.2. Case 2. Sample designs with pre-specified sample size, *n*

Consider the situation where the sample size (*n*), rather than the number of drivers required to be involved in rear-end crashes (*k*), is pre-specified and the interest is in estimating the expected number $\hat{k}$ of drivers who would be involved in rear-end crashes out of *n*. The proposed approach to resolve the sampling issues in this situation is demonstrated through examples in the following sections. GES and FARS data for the year 2000 were used to estimate different quantities required in estimating the sample sizes n_i and hence n ($= \sum_{i=1}^{M} n_i$). As in the previous example, the population of 18 to 64 year old drivers was stratified by

driver's age and that of 18 to 24 old drivers by driver's sex. The results were obtained using four methods of sample allocation: (i) n_i's determined by CIPI, (ii) equal n_i's, (iii) n_i's proportional to strata sizes, and (iv) n_i's adjusted using sampling strategy. The steps required for adjusting n_i's based on the sampling strategy are shown in Figure 4. The estimates of the

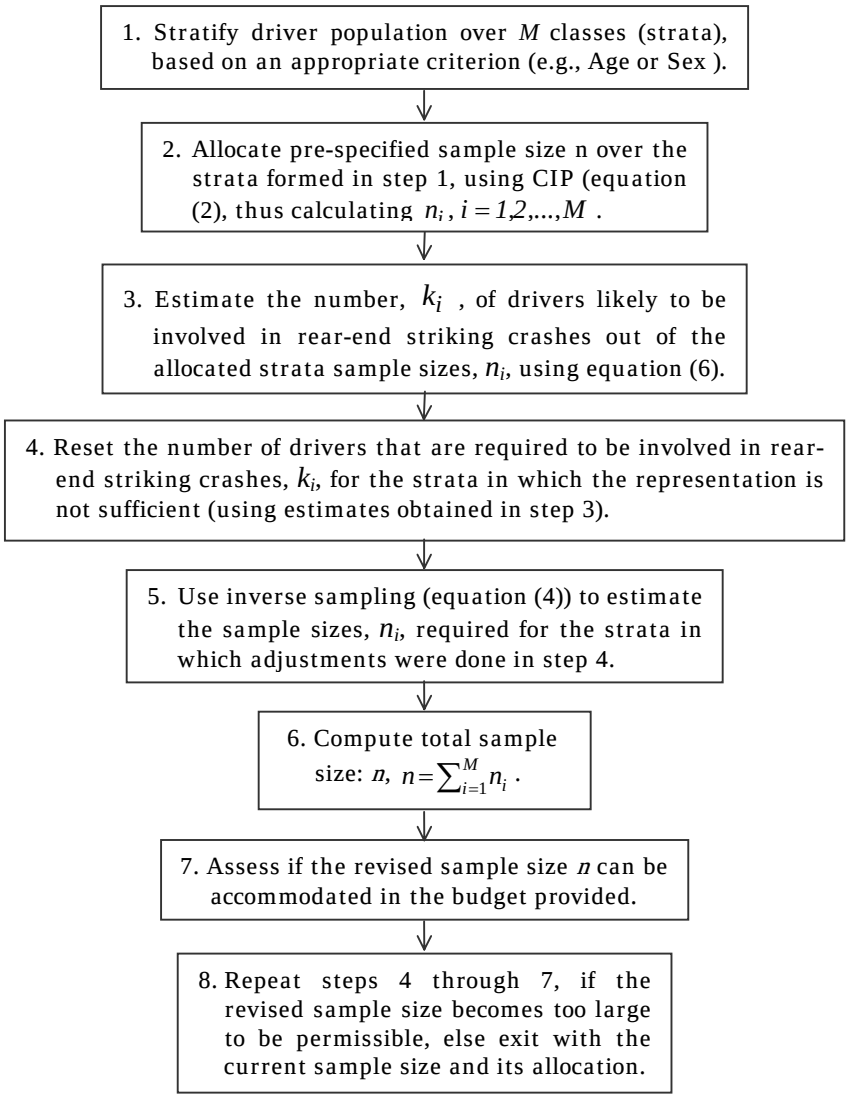

Figure 4. Sampling strategy with restriction on the sample size

strata sample sizes based on CIPI and the sampling strategy use inferences made in the previous sections as well as the estimates obtained there. While using the sampling strategy, it is assumed that some adjustment in the number of available experimental vehicles/drivers is permissible.

The results using the four methods of allocation mentioned above are presented in Table 3 (age-based stratification) and Table 4 (sex-based stratification). The last three columns (columns 5, 6, and 7) in these tables, respectively, represent the strata sample sizes (n_i) out of

the pre-specified sample size *n*, estimates of the corresponding numbers of drivers likely to be involved in rear-end crashes (k_i) and their 95% confidence intervals.

6.2.1. Example 1: Sample design with pre-specified sample size, *n*=200, in sampling from a population stratified by driver's age

Consider the situation where the population of 18 to 64 old drivers is partitioned on the criterion of age and the interest is in allocating *n* (=200) drivers in order to observe maximum possible number of them involved in rear-end crashes. As mentioned earlier, four methods of allocation were used. The results presented in Table 3 show that CIPI-based, equal, and proportional allocations would, respectively, result in 7, 6, and 4 drivers involved in rear-end crashes, out of the pre-specified sample of 200 drivers. On the other hand, the allocation based on the sampling strategy would require 7.5% increase in the total sample size in order to increase the representations of Age group 3 and Age group 4 among the crash-involved drivers. Specifically, with the sampling strategy, a sample of 215 drivers would be required in order to observe 7 drivers involved in rear-end crashes.

The sample allocation based on CIPI suggests that the selection of 150 drivers from Age group 2, 27 from Age group 3 would, respectively, result in 6 and 1 driver(s) involved in rear-end crashes. The point estimate of k_3 corresponding to Age group 4 in this allocation indicates that no driver would be involved in rear-end crash from this stratum. However, to be optimistic, as the upper confidence limit of the estimate of k_3 suggests, the selection of 23 drivers from this age group might result in the involvement of 1 driver in a rear-end crash.

The equal allocation of 200 drivers suggests that of the 6 drivers likely to be involved in rear-end crashes, 3 would be from Age group 2, 2 from Age group 3, and 1 from Age group 4. The proportional allocation, on the other hand, suggests a completely different sample design. The selection of 27, 99, and 74 drivers, respectively, from Age group 2, Age group 3, and Age group 4 is likely to result in 1, 2, and 1 driver(s) involved in rear-end crashes, respectively.

This example demonstrates that of the three allocations: CIPI-based, equal, and proportional, CIPI-based allocation is likely to generate data on the maximum number (7) of crash-involved drivers. The equal allocation can be considered as the second best choice that is likely to result in 6 crash-involved drivers. Due to poor performance (with the likelihood of only 4 drivers involved in rear-end crashes), the proportional allocation cannot be recommended in this case. In fact, the strata sizes resulting from this allocation are too heterogeneous and are not commensurate with the corresponding crash involvement propensities.

In a given situation, for certain reasons, the representation of some strata in the sample may not be satisfactory, thus requiring a revision of the sample allocation. The sampling strategy proposed in Figure 4 can help a great deal in resolving this issue. Using this strategy, the strata sample sizes were estimated as 100, 65, and 50 drivers, respectively, from Age group 2, Age group 3, and Age group 4. As compared to CIPI-based allocation, the revised allocation requires significant increase in the sample sizes from Age group 3 and Age group 4. This, however, is adjusted to a large degree by a decrease in the sample size from Age group 2, so as to keep the total sample size to its minimum possible. The results show that the revised

allocation would result in 4, 2, and 1 driver(s) involved in rear-end crashes, respectively, from Age group 2, Age group 3, and Age group 4.

Table 3. Sample designs with pre-specified sample size, *n*=200, using CIPI, equal and proportional allocations and the sampling strategy in sampling from a stratified (by age) population of drivers

Allocation criterion	Stratum of drivers (i)	Probability rear-end crash involvement (p_i)	Constant of proportionality (α_i)	Allocated stratum sample size ($n_i = \alpha_i \cdot n$)	Drivers likely to be involved in rear-end crashes (k_i)	95% Confidence interval [k_{iLOWER}, k_{iUPPER}]
CIPI	1. Age group 2	0.03742	0.74949	150	6	[1, 11]
	2. Age group 3	0.02393	0.13294	27	1	[0, 3]
	3. Age group 4	0.01575	0.11758	23	0	[0, 1]
	Total			200	7	
EQUAL	1. Age group 2	0.03742	0.33333	67	3	[0, 6]
	2. Age group 3	0.02393	0.33333	67	2	[0, 4]
	3. Age group 4	0.01575	0.33333	66	1	[0, 3]
	Total			200	6	
PROPORTIONAL	1. Age group 2	0.03742	0.13721	27	1	[0, 3]
	2. Age group 3	0.02393	0.49472	99	2	[0, 5]
	3. Age group 4	0.01575	0.36807	74	1	[0, 3]
	Total			200	4	
SAMPLING STRATEGY	1. Age group 2	0.03742	*n*	100	4	[0, 8]
	2. Age group 3	0.02393	*n*	65	2	[0, 4]
	3. Age group 4	0.01575	*n*	50	1	[0, 3
	Total			215	7	

Source: National Center for Statistics and Analysis, NHTSA, GES and FARS 2000, FHWA
n not required

6.2.2. Example 2: Sample design with pre-specified sample size, *n*=200, in sampling from a population stratified by driver's sex

The pre-specified sample size *n* (=200) was allocated over two sex-based strata of young drivers by using three methods of allocation: CIPI, equal, and proportional to the strata sizes. The results are presented in Table 4. These results show that the sample sizes assigned on the basis of CIPI, equal, and proportional allocations would result, respectively, in 8, 7, and 7 drivers involved in rear-end crashes.

Table 4. Sample designs with pre-specified sample size, *n*, using CIPI, equal and proportional allocations in sampling from a stratified (by sex) population of young drivers

Allocation criterion	Stratum of drivers (i)	Probability rear-end crash involvement (p_i)	Constant of proportionality (α_i)	Allocated stratum sample size ($n_i = \alpha_i \cdot n$)	Drivers involved in rear-end crashes (k_i)	95% Confidence interval [$k_{i\,LOWER}, k_{iUPPER}$]
CIPI	1. Young male	0.04199	0.55045	110	5	[1, 9]
	2. Young female	0.03260	0.44955	90	3	[0, 6]
	Total			200	8	
EQUAL	1. Young male	0.04199	0.50000	100	4	[0, 8]
	2. Young female	0.03260	0.50000	100	3	[0, 6]
	Total			200	7	
PROPORTIONAL	1. Young male	0.04199	0.51268	110	4	[0, 8]
	2. Young female	0.03260	0.48732	90	3	[0, 6]
	Total			200	7	

Source: National Center for Statistics and Analysis, NHTSA, GES and FARS 2000, FHWA

The CIPI-based allocation suggests the selection of 110 young male and 90 young female drivers in the sample that is likely to result in the involvement of 5 young male and 3 young female drivers in the rear-end crashes. The equal allocation suggests that a sample comprised of 100 young male and 100 young female drivers is likely to result in 4 young male and 3 young female drivers. Although the proportional allocation also suggests the involvement of 4 young male and 3 young female drivers in rear-end crashes, the required sample sizes are different. This allocation suggests the selection of 103 young male and 97 young female drivers to compose a sample of 200 young drivers.

This example demonstrates that highest number (8) of drivers are likely to be involved in rear-end crashes, if the sample of pre-specified size (200) is selected from the population of young drivers and CIPI is used as allocation criterion. The equal and proportional allocations are the other possible choices with slightly different sample allocation, but with the likelihood of 7 drivers involved in rear-end crashes.

7. Conclusions and recommendations

The contingency analysis of GES and FARS data for the year 2000 provided strong evidence of the association between driver attributes, age and sex, and the crash involvement of drivers (rear-end or otherwise). These factors can, therefore, be used for stratifying the population in order to have appropriate representation in the sample from age- or sex-based classes of drivers for data collection on rear-end crashes.

As demonstrated through examples, the statistic CIPI can provide a useful guideline to optimally allocate the sample over the strata by making greater provision in the sample for the strata that are more prone to rear-end crash involvement. Based on this statistic, it was found that 18 to 24 year old drivers were most prone to rear-end crash involvement. Thus, when age was used as the stratification criterion, out of 319 drivers required for 10 of them to be involved in rear-end crashes, the largest number (n_1 = 213) was allocated to this stratum. Similarly, among young drivers, male drivers were found to be more prone to rear-end crash involvement. Thus, when sex was used as the criterion for stratifying only the young drivers, a larger number (143) was allocated to the stratum of young male drivers as compared to 123 allocated to the stratum of young female drivers. When the sample size was assumed fixed (n =200) in advance and the population was stratified by age, the largest number, 150, was recommended from the stratum of young male drivers, with the likelihood of 6 drivers being involved in rear-end crashes.

The CIPI-based sample allocation in both modes of stratification, driver's age and sex, was compared with some other possible methods of allocation, equal number of drivers from each stratum and strata sample sizes proportional to the strata sizes. The examples illustrated that due to the differential that exists among the strata with respect to the crash involvement propensity, for the same target number of crash-involved drivers, both equal and proportional allocations resulted in larger strata sample sizes and hence larger total sample size as compared to the one suggested by CIPI-based allocation. Similarly, in the case of pre-specified sample size, both equal and proportional allocation, indicate the likelihood of a smaller number of crash-involved drivers as compared to CIPI-based sample allocation.

So far as the usefulness of the proposed sampling strategies is concerned, two types of situations can be handled: (i) when some strata are not well represented in the target number of rear-end crash-involved drivers, and (ii) when some strata sample sizes (out of pre-specified sample size) do not promise a satisfactory number of crash-involved drivers.

The proposed sampling strategies are neither data dependent nor population dependent. In fact, the approach used in this study is fairly general and can be used for resolving similar sampling issues involved in data collection in similar setups.

8. Appendix A. Contingency tables

Table A.1. This table shows sample size and weighted size for each class (age group), that were computed by SUDDAN while carrying out the contingency analysis for testing the association between *Age* and *Crash event* (Section 4.1).

Table A. 1. Contingency table: Driver's *Age* vs. *Crash event*

Driver's age	Statistics	Crash event		Total
		Rear-end	Other	
Age group 1 (< 18)	Sample size	1,708	4,097	5,805
	Weighted size	234,832	463,310	698,142
Age group 2 (18 to 24)	Sample size	6,243	13,731	19,974
	Weighted size	791,242	1,460,653	2,251,895
Age group 3 (25 to 44)	Sample size	15,162	27,471	42,633
	Weighted size	1,795,324	2,744,902	4,540,226
Age group 4 (45 to 64)	Sample size	7,852	14,325	22,177
	Weighted size	893,748	1,458,380	2,352,128
Age group 5 (>64)	Sample size	3,165	8,514	11,679
	Weighted size	422,690	1,052,585	1,475,275
Total	Sample size	34,130	68,138	102,268
	Weighted size	4,137,836	7,179,830	11,317,666

Source: National Center for Statistics and Analysis, NHTSA, GES and FARS 2000

Table A.2. This table shows sample size and weighted size for each class (*Sex* of young driver), that were computed by SUDDAN while carrying out the contingency analysis for testing the association between *Sex* of the young driver and *Crash event* (Section 4.1).

Table A.2. Contingency table: Young driver's *Sex* vs. *Crash event*

Driver Age/ Sex	Statistics	Crash event		Total
		Rear-end	Other	
Young (18 to 24) Male	Sample size	3,701	8,541	12,242
	Weighted size	453,724	878,006	1,331,730
Young (18 to 24) Female	Sample size	2,542	5,186	7,728
	Weighted size	337,517	582,519	920,036
Total	Sample size	6,243	13,727	19,970
	Weighted size	791,241	1,460,525	2,251,766

Source: National Center for Statistics and Analysis, NHTSA, GES and FARS2000

9. Appendix B. Analytical details of Crash Involvement Propensity Index

B.1 Crash Involvement Propensity Index

This appendix supplies the analytical details of the statistic *Crash Involvement Propensity Index* (CIPI), used in Section 4.2.

Consider a situation in which, based on a certain criterion, the drivers are divided into *M* subpopulations and our interest is in comparing these subpopulations with respect to their propensity of being involved in a rear-end crash. In order to develop a reasonable measure of the crash involvement propensity of a driver belonging to a subpopulation as compared to other subpopulations, it is important to consider the occurrence of rear-end crash-involved drivers in this subpopulation relative to the occurrence of its drivers in the entire population of drivers. The important information that one needs in this context is an answer to the question: Given that a driver selected at random is from a certain subpopulation, what is the probability that he/she would be involved in a rear-end crash? In other words, what one needs to look for is the *likelihood* of a driver of each subpopulation being involved in a rear-end crash.

For this purpose, we consider the space Ω of all drivers belonging to a subpopulation and the subspace Ω_C of those drivers of this subpopulation who are involved in rear-end crashes.

Let

$L(\Omega)$ be the probability that a driver selected at random belongs to the subpopulation Ω,

$L(\Omega_C)$ be the probability that a driver selected at random is involved in a rear-end crash, given that he/she belongs to the subpopulations Ω.

Let N be the number of drivers in the entire population of drivers that has been divided into M subpopulations, based on a certain criterion, S_i the number of drivers in the subpopulation i, and C_i the number of drivers who are involved in rear-end crashes from this subpopulation, $i = 1, 2, \ldots, M$. Then the crash involvement propensity of drivers belonging to subpopulation i can be defined as

$$\lambda_i = \frac{L(\Omega_C^{(i)})}{L(\Omega^{(i)})}, \quad i = 1, 2, \ldots, M, \tag{B.1}$$

where

$$L(\Omega_C^{(i)}) = \frac{C_i}{S_i}, \quad S_i > 0$$

and

$$L(\Omega^{(i)}) = \frac{S_i}{N},$$

so that λ_i defined in (B.1) becomes

$$\lambda_i = N\left(\frac{C_i}{S_i^2}\right), \quad i = 1, 2, \ldots, M \cdot \tag{B.2}$$

Note that λ_i in (B.2) is the conditional probability of a driver being involved in a rear-end crash, given that he/she belongs to that *i*-th subpopulation. In order to compare the crash involvement propensity of mutually disjoint subpopulations $A_1, A_2,\ldots,A_M$ into which the population of all licensed drivers is divided, these probabilities can be combined to define the *Crash Involvement Propensity Index* (CIPI)

$$\phi_i = \frac{\lambda_i}{\sum_{j=1}^{M} \lambda_j}, \qquad i = 1, 2, \ldots, M \cdot$$

Using λ_i from (B.2), the CIPI can be derived in the usable form

$$\phi_i = \frac{\frac{C_i}{S_i^2}}{\sum_{j=1}^{M}\left(\frac{C_j}{S_j^2}\right)}, \quad i = 1, 2, \ldots, M \cdot \tag{B.3}$$

10. References

[1] Wilks, S. S., *Mathematical Statistics,* John Wiley and Sons, New York (1962).

[2] Singh, Santokh, *Driver attributes and rear-end crash involvement propensity,* (NHTSA Technical report No. DOT HS 809 540, March 2003.

DOT HS 809 541
April 2003

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

People Saving People
www.nhtsa.dot.gov