A SAMPLING STRATEGY FOR REAR-END PRE-CRASH DATA COLLECTION

Santokh Singh

NCSA, National Highway Traffic Safety Administration
400 Seventh Street SW, Washington, DC 20590
Santokh.Singh@nhtsa.dot.gov

Abstract

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 the driver's perception of the complex scenario that emerges prior to the crash, but also on his/her pre-crash driving behavior, response to the imminent crash situation, and performance in resolving the driving conflicts. Obviously, any effort directed toward crash countermeasures must start from data collection in order to have a better understanding of these driver-related parameters in 'naturalistic' settings. This would require deploying vehicles on the roadways that are equipped with certain devices to record data on such parameters. Due to the random nature of these crashes, the number of vehicles actually required to obtain the desired amount of information may be large, thereby making data collection an expensive proposition. A sample design is needed that can conserve resources and yet obtains the maximum information.

The present study aims at proposing a sampling strategy for designing an optimal sample. It consists of stratifying the target population by driver's age and allocating the sample over the strata on the basis of driver's propensity of being in the striking/struck role in a rear-end crash. With a specific requirement of observing certain numbers of drivers in these two roles, the proposed strategy is compared with other methods of allocation, such as equal and proportional. The sample designed through this strategy is found to be much more economical in terms of the number of vehicles that need to be equipped and the number of voluntary drivers required.

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 2001. In terms of drivers crash involvement, of the 191,275,719 licensed drivers in 2001reported by Federal Highway Administration (FHWA), approximately 2.1% were involved in rear-end crashes, making up 36.5% of all drivers involved in various types of crashes. These figures suggest the necessity of developing crash countermeasures that could reduce the occurrence of 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 drivers on the roadways. The vehicles deployed for this type of data collection must therefore be equipped with certain devices that could record the data 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 designing an optimal sample for rear-end pre-crash data collection. A probabilistic approach is used to formulate a sampling strategy for this purpose.

2. Objective of the Study

In order to develop effective rear-end crash countermeasures and accurately estimate their potential benefits, greater understanding is needed regarding specific parameters of driving behavior and performance of drivers. These parameters are determined by the complex scenario that emerges prior to a rear-end crash. In this context it is important to remember that a driver involved in a rear-end crash can be the driver of a striking or struck vehicle, or even of the vehicle that strikes as well as struck by another vehicle. Therefore, information on the parameters related to both striking and struck drivers is useful for the development of countermeasures. While data collection is crucial for acquiring this information, 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, an important aspect of the data collection process is that the vehicles comprising the sample have to be kept deployed until the required numbers of drivers have been involved in rear-end striking and struck 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. The objective of the present study is to propose a sampling strategy that is optimal in terms of the content of information about the driver-related parameters across the population of drivers. A diligent selection of drivers from the target population is crucial for arriving at an optimal sample design. The present study is an attempt in this direction and is focused on

- Choosing a criterion that could be used to stratify the population
- Choosing a sample allocation criterion

These criteria will be used in the sampling strategy to design a sample with the pre-specified numbers of drivers that are required to be in the striking and struck roles in the rear-end crashes.

3. Data Sources and Variables for Analysis

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

- Age distribution of licensed drivers in 2001, reported by the FHWA
- Drivers involved in rear-end and other crashes in 2001, reported in GES
- Drivers involved in fatal crashes in 2001, 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.

The first and foremost task in designing a sample is to select the relevant factors (variables) from a large number of variables coded in GES and FARS databases. Keeping in view the fact that our interest is in the driving behavior and driver's performance prior to a rear-end crash, the factor (variable) that most deserves attention is the *Manner of collision* in the crash; we will focus on *Rear-end* crashes only. The other variables that need to be included in the context of the present study are the *Occupant role* and the *Vehicle role*; we will consider the occupant's role as *Driver*, while two types of *Vehicle role* will be considered; namely *Striking* and *Struck*. *Vehicle role: Both* will be included in *Striking* as well as in *Struck*.

Given a crash, the information about the manner of collision, occupant's role, and vehicle role can be combined by defining a new variable *Crash event*, which will be used in the subsequent analysis and discussion.

$$Crash \quad event = \begin{cases} \text{1, if the manner of collision is rear-end and the vehicle/driver role is } \textit{Striking}; \\ \text{2, if the manner of collision is rear-end and the vehicle/driver role is } \textit{Struck}; \\ \text{3, if the driver is involved in a crash other than the rear-end or is not involved in any type of crash.} \end{cases}$$

Last but not least, the perception of the circumstances surrounding a crash as well as the driving behavior and performance of the driver prior to a crash, seem to be related to driver's age. The related variable, coded as *Age* in GES and FARS databases, will be the basis of all analysis done in this study.

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, achieving this aim is easier if the target population is stratified using an appropriate criterion. An efficient criterion to allocate the sample over the strata is an additional tool in designing an optimal sample.

4.1. Stratification Criterion

As mentioned earlier, the involvement of a driver in a crash depends on his/her perception of the complex scenario that emerges prior to a crash and that the driver's pre-crash driving behavior and performance play important roles in resolving the driving conflicts. This suggests that driver's age may be one of several factors contributing to the rear-end crash involvement of a driver and the role he/she assumes in the crash. This attribute can therefore be considered as one of the possible factors for stratification of the target population. Nevertheless, using this factor for this purpose will make sense only if there is an evidence of the association between *Age* and *Crash event* (the variable that defines a driver's role in a rear-end crash).

Contingency analysis was performed for testing the association between *Age* and *Crash event*. In order to test the independence between these categorical variables, the drivers were classified using the following mode:

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 classification in place, the contingency analysis of GES data for the year 2001 was carried out to test the hypothesis: *there is no association between 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 test procedure yields χ^2 = 279.9 with 8 degrees of freedom. The 95th percentile 9.49 of χ^2 distribution (with 8 degrees of freedom) being far less than 279.9, the hypothesis of no association is discredited, thereby indicating that there is a strong evidence of an association between *Age* and *Crash event*. This inference provides a strong reason to use driver's age as the stratification criterion in designing the sample. In the subsequent discussion, the classes of drivers defined on the criterion of *Age* will be referred to as the attribute-based classes.

4.2. Sample Allocation Criterion

As mentioned earlier, the basic aim is to design a sample so that a maximum number of drivers involved in rear-end crashes may be observed with minimum number of vehicles deployed. The stratification based on age can be effectively used in this 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 *j*th stratum being in one of the two roles 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 from a stratum in the striking/struck role 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 and assume one of the two roles? These probabilities can then be combined into the statistic ϕ_j, called Crash Involvement Propensity Index (CIPI) (see details in Appendix A) as proposed in [2]

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

where

C_j is the number of rear-end crash-involved drivers belonging to the *j*th stratum (*j*th subpopulation);

S_j is the number of drivers in the *j*th stratum, i.e., the size of the *j*th subpopulation; $S_j > 0$, $S_1 + S_2 + \ldots + S_M$ with N_T as the size of the population of all drivers;

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

Note that the numerator in (1) takes into account the likelihood or conditional probability (conditional on stratum) of a driver belonging to the *j*th stratum being involved in a rear-end crash, while the denominator is the normalizing quantity. Obviously, the statistic ϕ_j satisfies the inequality $0 \le \phi_j \le 1$. The statistic CIPI given in equation (1) provides a measure of the propensity of drivers belonging to a certain stratum of being in the striking/struck role in the rear-end crashes, relative to that of the drivers of other strata.

It is obvious that more drivers are expected to be involved in rear-end striking (struck) crashes from a stratum in which drivers have higher propensity of being in the striking (struck) role in rear-end crashes. This index can therefore be used as the constant of proportionality for optimally allocating the sample over the strata. Following this argument, the numbers k_1 (striking) and k_2 (struck) can be disbursed over the strata using the relation

$$k_{ij} = k_i\ \phi_j\ ,\quad i = 1,\ 2;\quad j = 1,\ 2,\ \ldots,\ M\ , \tag{2}$$

where ϕ_j is given by (1) and k_{1j} and k_{2j} are, respectively, the required numbers of drivers in the striking and struck roles from the *j*th stratum. Subsequently, using inverse sampling (to be discussed latter in the discussion), the sub-sample size in each stratum and hence the total sample size can be determined.

4.3. Guidelines from the Crash Involvement Propensity

Before CIPI is used for providing guidelines for designing the sample, i.e., in stratifying the population of drivers, it is important to remember that the sample is supposed to consist of voluntary drivers. In that case, due to the anticipated operational difficulties, Age group 1 and Age group 5 should preferably be excluded from the sample. Accordingly, the target population considered in this study consists of 18 to 64 year-old drivers (Age group 2, Age group 3, and Age group 4). The allocation criterion proposed in (1) was calculated for three age groups of drivers as decided earlier, using GES and FARS data for the year 2001. The results on the crash involvement propensity (equation (1)) presented in Figure 1 show that Age group 2 (18 to 24) drivers have the highest propensity of being in the striking and struck roles as compared to 25 to 64 year-old drivers, who have much lower propensity of assuming any of these roles. This shows that, when engaged in data collection, many more drivers can be expected to have been involved in the rear-end crashes both as striking and struck drivers from Age group 2. These results also show that, when involved in a rear-end crash, a 25 to 64 year-old driver is more likely to be in the striking role than in the struck, though this tendency is higher among Age group 3 drivers as compared to Age group 4 drivers. The tendencies of drivers of different age groups in assuming a role in a rear-end crash that this statistic indicates can be exploited in designing a sample.

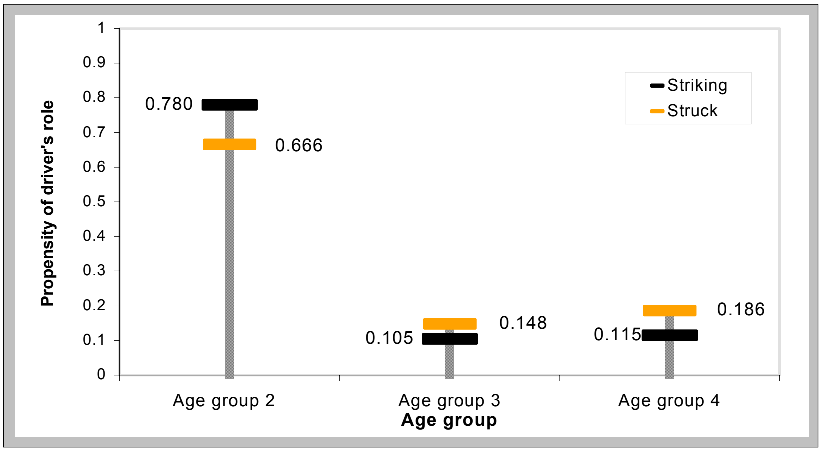

Figure 1. Propensity of being in the striking and struck role in a rear-end crashes three age groups.

5. Sampling Issue in Rear-End Pre-Crash Data Collection

Although rear-end crashes form a significant proportion of crashes, their frequency is not so high that a significantly large amount of data can be generated by engaging only a small number of experimental units. Being specific about the number of observations required on the striking and struck drivers is, therefore, important for the success of any data collection project with the current theme.

Consider the situation where the aim is to obtain information on the driver-related parameters for k_1 striking and k_2 struck drivers. Then the sampling issue involved is to estimate the number, *N*, of vehicles to be deployed for data collection (i.e., the sample size) that is large enough to include at least k_1 striking and at least k_2 struck drivers.

6. Probabilistic Formulation of the Sampling Problem

Like all other road events, the occurrence of rear-end crashes and the role that a driver is likely to assume in such crashes are random events. Accordingly, the variable *Crash event* defined earlier is a discrete random variable. This further means that the phenomenon of drivers' involvement in rear-end crashes and the roles that they assume therein can be treated probabilistically.

6.1. Categorization of Drivers in the Target Population

Consider the target population of drivers (i.e., 18 to 64 year old drivers). Based on the variable *Crash event* defined in Section 3, the drivers in this population can be categorized into the following categories:

C_1: drivers involved in the rear-end crashes with their vehicles in the striking role;
C_2: drivers involved in the rear-end crashes with their vehicles in the struck role;
C_3: drivers involved in crashes other than the rear-end crashes or not involved in any crash.

This event-based categorization of drivers classified into age groups A_2, A_3, and A_4 is shown in Figure 2. Note that the areas demarked in Figure 2 are merely representative of the class/category of drivers and not of their actual sizes. In the subsequent analysis, the age-based classes A_2, A_3, and A_4 will be referred to also as subpopulations. Having done this categorization of each subpopulation, the task is to determine the sub-sample sizes.

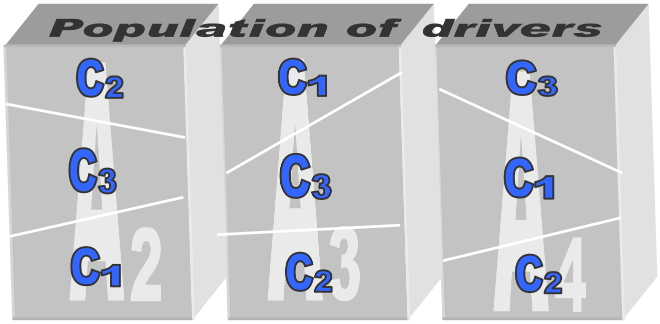

Figure 2. Categorization of drivers (classified by Age), based on Crash event.

For that purpose, define events E_1, E_2, and E_3 as

E_1: driver is involved in a rear-end crash and is the driver of the striking vehicle;
E_2: driver is involved in a rear-end crash and is the driver of the struck vehicle;
E_3: 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 three categories C_1, C_2, and C_3, depending on the occurrence of the events E_1, E_2, and E_3, respectively. For instance, if a driver is involved in a rear-end crash and is the driver of the striking vehicle, 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_3. This categorization of drivers will be referred to as the event-based categorization.

Due to the uncertainty inherent with the road events, each of the event-based categories of subpopulation can be associated with certain probability. Specifically, let

p_1 = P{E_1}, i.e., the probability that a driver belongs to category C_1 of the population;

p_2 = P{E_2}, i.e., the probability that a driver belongs to category C_2 of the population;

p_3 = P{E_3}, i.e., the probability that a driver belongs to category C_3 of the population ($p_3 = 1 - p_1 - p_2$).

6.2. Inverse Sampling

Recall that the objective of this study is to design a sample to collect data on at least k_1 drivers involved in rear-end crashes assuming the role of striking drivers and k_2 involved in such crashes assuming the role of struck drivers. This further means that one needs to continue collecting data on the pre-crash driving behavior and performance of drivers until at least the specified numbers of drivers k_1 and k_2, respectively, have fallen in the categories C_1 and C_2. This immediately suggests that the process of data collection can be thought of as throwing balls successively into three boxes, until at least k_1 balls of C_1-type type are obtained in the first box and at least k_2 balls of C_2-type are obtained in the second box. Since the outcome of a throw resulting in a ball that goes into the third box does not affect termination of this process, no attention is paid to what happens to that box. Thus, the sample size in this experimentation (throwing balls into three boxes) is nothing but the number of trials one has to wait through in order to achieve the objective. Correspondingly, speaking in terms of drivers/vehicles, the sample size in the present context is the number of drivers/vehicles one has to wait through in order to collect data on at least k_1 drivers in the striking role and at least k_2 drivers in struck role. In statistical terms, this process is called 'inverse sampling' and is described in the flow chart presented in Figure 3, where N_s is the number of striking drivers, and N_{st} the number of struck drivers.

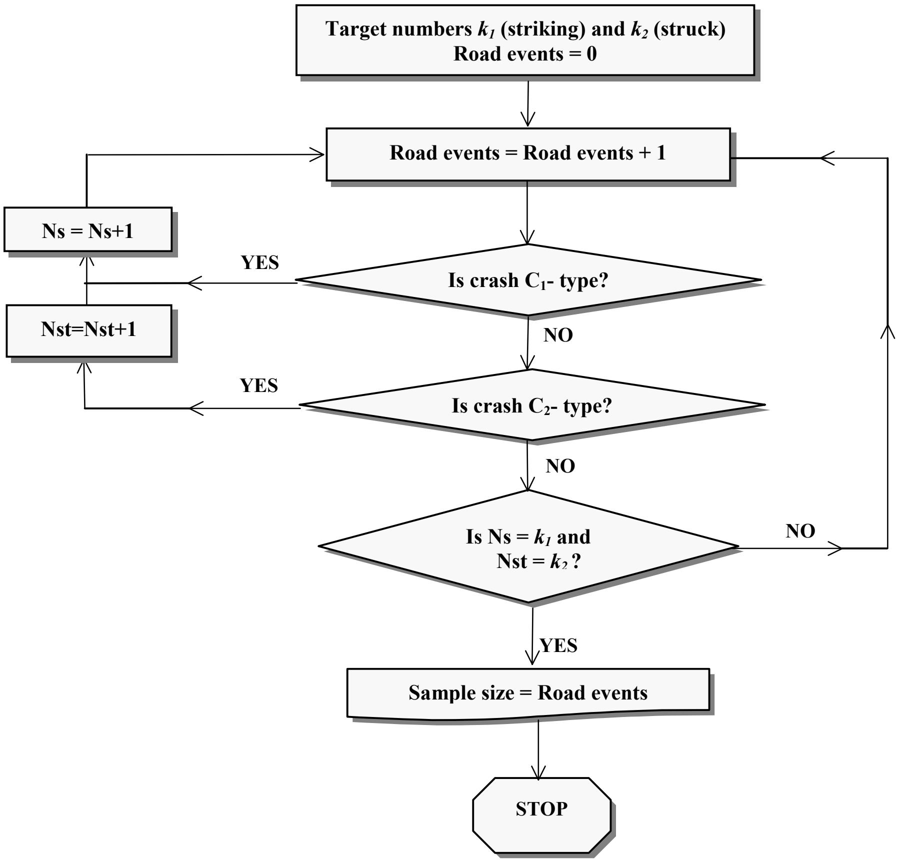
Target numbers k_1 (striking) and k_2 (struck)
Road events = 0
Road events = Road events + 1
Ns = Ns+1
YES
Is crash C_1- type?
Nst=Nst+1
NO
YES
Is crash C_2- type?
NO
Is Ns = k_1 and Nst = k_2?
NO
YES
Sample size = Road events
STOP

Figure 3. Flow chart describing inverse sampling.

A typical sequence of the drivers involved in all types of road events, before 3 of them are involved in C_1-type of crashes and 2 in C_2-type, would look like shown in Figure 4.

<table>
<tr><th>Driver/Vehicle</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>11</th><th>12</th><th>13</th><th>14</th><th>15</th></tr>
<tr><td>Crash Event</td><td>C_3</td><td>C_2</td><td>C_3</td><td>C_3</td><td>C_3</td><td>C_1</td><td>C_3</td><td>C_3</td><td>C_2</td><td>C_3</td><td>C_1</td><td>C_2</td><td>C2</td><td>C_3</td><td>C_1</td></tr>
<tr><td>Run of</td><td colspan="2">2</td><td colspan="4">4</td><td colspan="3">3</td><td colspan="2">2</td><td>1</td><td>1</td><td colspan="2">2</td></tr>
</table>

Figure 4. Sequence of road events before at least 3 drivers are involved in rear-end striking and 2 in rear-end struck crashes.

It would not be out of context to mention at this point that the total number of drivers, 15 in this instance, is subject to randomness due to uncertainty involved in road events.

6.3. Probability Distribution of the Number of Striking and Struck Drivers Involved in Rear-End Crashes

It is now the time to put all the information together which we gained about rear-end crashes. First of all, it is clear that as the experimentation (data collection) goes on, the vehicles deployed on the road and hence the drivers engaged would be involved in crashes; the probability being p_1 that a driver involved in a rear-end crash would be the *Striking* driver and p_2 the probability that a driver involved in a rear-end crash would be the driver who was *Struck*. Any other type of road event, not covered by the above two crash scenarios, occurs with probability p_3 (=1 - p_1 - p_2). It is important to note that the termination of data collection depends only on the number of drivers assuming striking and struck roles that have happened up to and including a rear-end crash and not on the occurrence of the other type of road event. However, the total number of drivers we will have to wait through (length of the sequence in Fig. 4) to terminate data collection depends on how often C_3 occurs. These facts provide sufficient reason to treat the sample size problem as a *discrete waiting-time* problem [3].

Define the event E as

$$E = \{\text{at least } k_1 \text{ drivers are involved in } C_1\text{-type of crashes and at least } k_2 \text{ drivers in } C_2\text{-type}\}$$

It is easy to see that the event E can happen in one of the two ways; namely when a crash occurs that ends with exactly k_1 drivers involved in C_1-type of crashes and at least k_2 drivers in C_2-type, or when a crash occurs that ends with exactly k_2 drivers involved in C_2-type of crashes and at least k_1 drivers in C_1-type. Although the happening of the event E is decided by the occurrence of C_1 or C_2-type of crash event, as mentioned earlier, the occurrence of all other road events included in C_3-type decides the length of the sequence of road events (all types) that would be required to have happened, before the event E happens. One such sequence is shown in Figure 4. Continuing the argument, it can be concluded that the number of drivers, N, required for the happening of E is a random variable that assumes values

$$k_1 + k_2,\ k_1 + k_2 + 1,\ k_1 + k_2 + 2, \ldots\ldots\ldots$$

and follows a *negative multinomial distribution* (see Appendix B). Accordingly, the probability that n vehicles/drivers will be required for the event E to happen is given by (Appendix B).

$$P\{N = n\} = \sum_{i=1}^{2} \binom{n-1}{k_i - 1} p_i^{k_i}\, \Sigma\, \frac{(n-k_i)!}{r_l\,!\; r_3\,!}\, p_l^{r_l}\, p_3^{r_3} \quad , \tag{3}$$

where the summation Σ is taken over r_l and r_3 $(l \neq i)$, such that, $r_l \geq k_l$ and $r_l + r_3 = n - k_i$.

7. Estimation of Sample Size

Once the probability distribution of N has been established, the sample size is nothing but the expected value of N. Using the probability distribution given by (3), the expected number of vehicles required in order to compose a sample that consists of at least k_1 drivers involved in C_1-type of crashes and at least k_2 drivers in C_2-type is given by (Appendix B)

$$E_N = \sum_{i=1}^{2} \left\{ \sum_{n=k_1+k_2}^{\infty} n \binom{n-1}{k_i - 1} p_i^{k_i}\, \Sigma\, \frac{(n-k_i)!}{r_l\,!\; r_3\,!}\, p_l^{r_l}\, p_3^{r_3} \right\} \tag{4}$$

where the summation Σ is taken over r_l and r_3 $(l \neq i)$, such that, $r_l \geq k_l$ and $r_l + r_3 = n - k_i$. For small values of p_1, p_2, p_3, the convergence of the sum in (4) is bound to be slow. In order to overcome this computational difficulty a simplified form of E_N can be derived (see Appendix B)

$$\hat{N} = \left[\frac{k_1}{p_1} \left\{ 1 - I\left(k_1 + 1, k_2; \frac{p_1}{p_1 + p_2} \right) \right\} + \frac{k_1}{p_2} \left\{ 1 - I\left(k_2 + 1, k_2; \frac{p_2}{p_1 + p_2} \right) \right\} \right], \tag{5}$$

where $I(l, m; a)$ is the Incomplete beta function with $l, m \geq 0$ and $0 \leq a \leq 1$ (Appendix B).

8. Example: Sample Design for Rear-End Pre-Crash Data Collection with the Objective of 10 Striking and 10 Struck Drivers

In this section, the proposed sampling strategy is implemented and compared with other strategies (equal and proportional allocation). GES and FARS data for the year 2001 were used to estimate the sample size required for observing k_1 (=10) drivers in the striking role and k_2 (=10) drivers in the struck role in the rear-end crashes. As suggested by the contingency analysis, age of the driver was used as the stratification criterion. However, as mentioned earlier, due to the anticipated operational difficulties, Age group 1 and Age group 5 were excluded in the following analysis. As a first step, the target numbers k_1 and k_2 were partitioned into 3 numbers each, k_{i1}, k_{i2}, k_{i3} + k_{i1} + k_{i2} = k_i, i= 1, 2,with k_{1j} representing the number of striking drivers and k_{2j} representing the number of struck drivers from stratum j. This was done in three ways: (i) k_{ij}'s determined by the propensity statistic given by (1), (ii) equal k_{ij}'s, and (iii) ki_j's proportional to strata sizes.

The results for case (i) are presented in Table 1. These results show that, if CIPI is used as the sample allocation criterion, then, of the 10 striking drivers involved in rear-end crashes, 8 would be from Age group 2, and 1 each from Age group 3 and Age group 4. Similarly, of the 10 struck drivers involved in rear-end crashes, 7 would be from Age group 2, 1 from Age group 3, and 2 from Age group 4. Using the estimate of the sample size from (5) for each group of the target population, the respective sub-sample sizes for the three age groups were found to be 410, 138, and 175. Thus, with age as the stratification criterion and CIPI as the allocation criterion, the expected total number of vehicles/drivers required for data collection adds up to 723.

Table 1. Sample design for observing 10 drivers in the striking role and 10 in the struck role, in rear-end crashes (stratification by age, allocation by crash involvement propensity).

Stratum of drivers	**REAR-END CRASHES: STRIKING DRIVERS**			**REAR-END CRASHES: STRUCK DRIVERS**			**Stratum sample size**
(j)	**Constant of Proportionality** (α_j)	**Number of Striking drivers** $(k_{1j} = \alpha_j.k_1)$	**Probability (Striking)** (p_1)	**Constant of Proportionality** (β_j)	**Number of Struck drivers** $(k_{2j} = \beta_j.k_2)$	**Probability (Struck)** (p_2)	(n_j)
1. Age group 2	0.780278	8	0.026936	0.665939	7	0.017725	410
2. Age group 3	0.104673	1	0.012703	0.148212	1	0.013868	138
3. Age group 4	0.11505	1	0.010875	0.1858484	2	0.013545	175
Total sample size required for 10 drivers in striking role and 10 drivers in struck role							723

Source: National Center for Statistics and Analysis, NHTSA, GES and FARS 2001, FHWA

For the purpose of comparison, two other methods of allocation were also considered: equal and proportional. The results presented in Table 2 show that with the numbers k_{i1} = 4, k_{i2} = 3, and k_{i2} = 3, i = 1, 2, larger number of drivers, 297 and 329, 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 drivers have a lower propensity of being in the striking/struck role as compared with Age group 2 drivers and yet the samples selected from them are supposed to produce almost the same number in these two roles. This in turn raises the requirement of total number of vehicles/drivers in the sample to a larger number (873) as compared to CIPI-based allocation that requires 723 drivers with the same outcome.

Table 2. Sample design for observing 10 drivers in the striking role and 10 in the struck role in rear-end crashes (stratification by age, equal allocation).

Stratum of Drivers	REAR-END CRASHES: STRIKING DRIVERS			REAR-END CRASHES: STRUCK DRIVERS			Stratum Sample Size
(j)	Constant of Proportionality (α_j)	Number of Striking Drivers ($k_{1j} = \alpha_j.k_1$)	Probability (Striking) (p_1)	Constant of Proportionality (β_j)	Number of Struck Drivers ($k_{2j} = \beta_j.k_2$)	Probability (Struck) (p_2)	(n_j)
1. Age group 2	0.4	4	0.026936	0.4	4	0.017725	247
2. Age group 3	0.3	3	0.012703	0.3	3	0.013868	297
3. Age group 4	0.3	3	0.010875	0.3	3	0.013545	329
Total sample size required for 10 drivers in striking role and 10 drivers in struck role							873

Source: National Center for Statistics and Analysis, NHTSA, GES and FARS 2001, FHWA

The target numbers, 10 of rear-end crash-involved drivers in striking role and 10 in the struck role, were disbursed also using proportional allocation (i.e., proportional to strata sizes). The results presented in Table 3 show that with this allocation, from Age group 2 only 1 driver is expected to be in the striking role and 1 in the struck. This would require 94 drivers from this age group to be included in the sample. In order to observe 4 striking and 4 struck drivers, 458 drivers need to be included from Age group 3. Similarly, in order to observe 5 striking and 4 struck drivers, 405 drivers need to be included from Age group 4. Thus, with proportional allocation, 957 vehicles/drivers would be required for the same target numbers of striking and struck drivers. This number is larger than the one suggested by equal allocation (873) and much larger as compared to sample size (723) suggested by CIPI-based allocation.

Table 3. Sample design for observing 10 drivers in the striking role and 10 in the struck role in rear-end crashes (stratification by age, proportional allocation).

Stratum of Drivers	REAR-END CRASHES: STRIKING DRIVERS			REAR-END CRASHES: STRUCK DRIVERS			Stratum Sample Size
(j)	Constant of Proportionality (α_j)	Number of Striking Drivers ($k_{1j} = \alpha_j.k_1$)	Probability (Striking) (p_1)	Constant of Proportionality (β_j)	Number of Struck Drivers ($k_{2j} = \beta_j.k_2$)	Probability (Struck) (p_2)	(n_j)
1. Age group 2	0.137863	1	0.026936	0.137863	1	0.017725	94
2. Age group 3	0.484651	5	0.012703	0.484651	5	0.013868	458
3. Age group 4	0.377486	4	0.010875	0.377486	4	0.013545	405
Total sample size required for 10 drivers in striking role and 10 drivers in struck role							957

Source: National Center for Statistics and Analysis, NHTSA, GES and FARS 2001, FHWA

9. Conclusions and Recommendations

The contingency analysis of GES and FARS data for the year 2001 provided strong evidence of the association between driver age and the crash involvement of drivers (rear-end or otherwise). Driver's age can therefore be used as a criterion for stratifying the population so that different age groups can have appropriate representation in the sample.

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. In order to produce maximum amount of data, this age group should therefore contribute most to the sample.

The CIPI-based sample allocation 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 example demonstrated 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.

The proposed sampling strategy is neither data dependent nor population dependent. In fact, the approach used in this study is general and can be used for designing an optimal sample for data collection in similar setups. The sample allocation criterion proposed in this study can also be used when the total sample size is fixed in advance and the requirement is merely to allocate the sample size over the strata in order to get the best out of the restricted sample size. In that case the fixed sample size needs to be disbursed using CIPI. The expected number of drivers involved in rear-end striking and struck cashes in each age group can then be obtained using the binomial distribution.

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.

[3] McCarthy, P. J., 'Approximate Solutions for Means and Variances in a Certain Class of Box Problems', Annals of Mathematical Statistics, Vol. 18 (3), 1947.

11. Appendix A. Analytical Details of Crash Involvement Propensity Index

11.1 Crash Involvement Propensity Index

This appendix supplies the analytical details of the statistic 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,...,M.* Then the crash involvement propensity of drivers belonging to subpopulation *i* can be defined as

$$\lambda_i = \frac{L(\Omega_C^{(i)})}{L(\Omega^{(i)})}, \; i = 1, 2, \ldots, M \, , \qquad \text{(A.1)}$$

where

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

and

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

so that λ_i defined in (A.1) becomes

$$\lambda_i = N\left(\frac{C_i}{S_i^2}\right), \; i = 1, 2, \ldots, M \, . \qquad \text{(A.2)}$$

Note that λ_i in (A.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,…,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)

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

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

$$\phi_i = \frac{\dfrac{C_i}{S_i^2}}{\sum_{j=1}^{M}\left(\dfrac{C_j}{S_j^2}\right)} , \quad i = 1,\ 2,\ .\ .\ .\ ,\ M\ . \qquad \text{(A.3)}$$

12. Appendix B. Analytical Details of the Probability Distribution and Expected Value of N

12.1. Probability Distribution of N

This appendix supplies the analytical details of the probability distribution of the random variable N, referred to in Section 6.3.

Consider the situation in which we are interested in the occurrence of crashes that involve ***at least*** k_1 drivers in C_1-type crash events (that occur with probability p_1) and ***at least*** k_2 drivers in C_2-type crash vents (that occur with probability p_2), while all other type of crash events (forming one category) occur with non-zero probability, say, p_3 (=1 - p_1 - p_2).. The sampling problem in this case can be visualized as throwing balls into three boxes one by one (each box representing a category of drivers C_1, C_2, or C_3) until at least k_1 balls are obtained in the first box and at least k_2 balls are obtained in the second box. No attention is paid to what happens to the third box, as the number of balls in this box is not decisive of the termination of experiment (throwing balls in the boxes). The number of balls required or the number of trials, *N,* one has to wait through in order to accomplish this task is a random variable which has a *Discrete waiting-time distribution*. Accordingly, the probability that at the n-th throw k_1 balls will be obtained in the first box, before k_2 balls are obtained in the second box without paying attention to what happens to the third box is immediately seen to be

$$\binom{n-1}{k_1-1} p_1^{k_1} \Sigma_1 \frac{(n-k_1)!}{r_2\,!\; r_3\,!} p_2^{r_2} p_3^{r_3} \ , \qquad \text{(B.1)}$$

where the summation Σ_1 is taken over r_2 and r_3, such that $r_2 \geq k_2$ and $r_2 + r_3 = n - k_1$.

Similarly, the probability that at the n-th throw k_2 balls will be obtained in the second box, before k_1 balls are obtained in the first box, without paying attention to what happens to the third box is immediately seen to be

$$\binom{n-1}{k_2-1} p_2^{k_2} \Sigma_2 \frac{(n-k_2)!}{r_1\,!\; r_3\,!} p_1^{r_1} p_3^{r_3} \ , \qquad \text{(B.2)}$$

where the summation Σ_2 is taken over r_1 and r_3, such that $r_1 \geq k_1$ and $r_1 + r_3 = n - k_2$.

Combining (B.1) and (B.2), the probability that n throws will result into k_1 balls of C$_1$-type in the first box and at least k_2 balls of C$_2$-type in the second box is given by

$$P(N=n) = \binom{n-1}{k_1-1} p_1^{k_1} \Sigma_1 \frac{(n-k_1)!}{r_2\,!\; r_3\,!} p_2^{r_2} p_3^{r_3} + \binom{n-1}{k_2-1} p_2^{k_2} \Sigma_2 \frac{(n-k_2)!}{r_1\,!\; r_3\,!} p_1^{r_1} p_3^{r_3} \qquad \text{(B.3)}$$

where the summation Σ_1 is taken over r_2 and r_3, such that $r_2 \geq k_2$ and $r_2 + r_3 = n - k_1$, and the summation Σ_2 is taken over r_1 and r_3, such that $r_1 \geq k_1$ and $r_1 + r_3 = n - k_2$.

12.2. Expected Value of N

This appendix supplies the analytical details of the expected value of the random variable N, used in Section 7.

Using (B.3), the expected number of trials required for the occurrence of at least k_1 balls of C$_1$-type and k_2 balls of C$_2$-type in the sample drawn from a population that is categorized into three categories is given by

$$E(N) = \sum_{n=k_1+k_2}^{\infty} \left\{ n\binom{n-1}{k_1-1} p_1^{k_1}\, \Sigma_1 \frac{(n-k_1)\,!}{r_2\,!\, r_3\,!} p_2^{r_2} p_3^{r_3} + n\binom{n-1}{k_2-1} p_2^{k_2}\, \Sigma_2 \frac{(n-k_2)\,!}{r_1\,!\, r_3\,!} p_1^{r_1} p_3^{r_3} \right\} \qquad \text{(B.4)}$$

The convergence of the sums involved in (B.4) depends on the magnitude of the probabilities, p_1 and p_2; the smaller the magnitude of these quantities, the longer it will take for the summation to converge. However, the problem can be reduced by one dimension if we consider only those balls that go into first two boxes, that is, by considering the first two boxes as one unit. Thus, instead of counting the number of balls going into the first two boxes separately, only the number of balls going in the single unit are counted; the probability being $p_1 + p_2$ of balls going into this unit. With this formulation, it is obvious that the number of balls necessary to obtain k_1 balls in the first box and k_2 balls in the second box of this unit is a random variable X which takes on values $k_1 + k_2$, $k_1 + k_2 + 1$, $k_1 + k_2 + 2$, ……… ∞ and follows a negative multinomial distribution.

The expected number of trials for the occurrence of the desired event is given by

$$E[N] = E[E[X(p_1 + p_2),\ \infty(p_3)]]$$

$$= E\left[E\left[k_1\left(\frac{p_1}{p_1+p_2}\right),\ k_2\left(\frac{p_2}{p_1+p_2}\right)\right](p_1+p_2),\ \infty(p_3)\right]$$

This can be further simplified as

$$E[N] = \frac{k_1}{p_1}\left\{1 - I(k_1+1,\ k_2,\ \frac{p_1}{p_1+p_2}\right\} + \frac{k_2}{p_2}\left\{1 - I(k_2+1, k_2,\ \frac{p_2}{p_1+p_2}\right\} \qquad \text{(B.5)}$$

where $I(k_1+1, k_2; p_1)$ is the value of the incomplete beta function which for the specified parameters l, m, and a is defined as

$$I(l, m; a) = \left\{\frac{\Gamma(l+m)}{\Gamma(l)\Gamma(m)}\right\}\int_0^a x^{l-1}(1-x)^{m-1}\,dx\,, \qquad \text{(B.6)}$$

where $l, m \geq 0$, $0 \leq a \leq 1$ and $\Gamma(m)$ is the value of the gamma function evaluated at m.