Attachment H: Sample Size and Sample Plan

Sample Strategy

A two-stage sampling approach will be used in the National Survey of U.S. Long-Haul Truck Driver Safety and Health. The first stage will be selection of truck stops where survey administration will occur, and the second stage will be selection of truck drivers at each truck stop to whom the survey will be administered. This is the most efficient and effective way of obtaining a representative sample of drivers for administration. In order to ensure that long-haul truck drivers stopping at truck stops along both heavily-traveled routes and lesser-traveled ones will be represented in the survey sample, a sample strategy has been developed in which a randomly selected sample of drivers stopping at each type of truck stop will be included. Survey administration including interview administration and anthropometric measurements will be done during personal interviews of selected drivers.

An optimal sample design balances survey cost factors with estimation procedures in order to obtain estimators with minimum variance at lowest cost. It is generally most cost efficient to plan to have the interviewers spend a fixed amount of time at each site. This allows for scheduling site visits in a straightforward way. Our sample design includes a fixed number of truck drivers *m* to be interviewed at each of *n* sampled truck stops. The *n* sampled truck stops will be sampled with probability proportional to estimated size (pps), where estimated size is the number of paved parking spaces for trucks at the truck stop.

Sample Size

In the National Survey of U.S. Long-Haul Truck Driver Injury and Health, long-haul truck drivers stopping at selected truck stops will be interviewed. Interviews will be administered at each truck stop for 3 days. Since the number of interviews to be completed at each truck stop during the data collection period is not expected to be the same and will depend on the traffic flow through that truck stop, truck stops are defined as 'high-flow' and 'low-flow' truck stops. High- or low-flow truck stops are determined by the truck stop's location on traffic corridors with truck traffic flows of 10,000 or more trucks per day, based on the Freight Analysis Framework Version 2 [Federal Highway Administration 2007]. The Freight Analysis Framework (FAF) integrates data from a variety of sources to estimate commodity flows and related freight transportation activity among states, regions, and major international gateways between 2002 and 2035. Because of the lower truck traffic flow in the low-flow truck stops, we assume that only half as many interviews will be completed in each low-flow truck stop as are completed in a high-flow truck stop. Under these assumptions, the best design is to have roughly 80% of the truck stops as high-flow and 20% low-flow, resulting in 90% of all interviews being done in high-flow truck-stops and 10% of all interviews being done at low-flow truck-stops. Travel costs to low-flow sites are assumed to be twice those for high-flow sites.

The number of truck drivers *m* to be interviewed at each of the *n* sampled truck stops was determined by minimizing the variance of estimator $\hat{Y}$ of characteristic Y under the sample design. Considering that the truck stops are stratified into high/low-flow strata, and accounting for

both between-site variability and within-site variability for each type of truck stop, the total variance for an estimator in this two stage sample would be [Levy and Lemeshow 1980]:

$$\sigma^2(Y) = \sigma^2_h(b) + \sigma^2_h(w) + \sigma^2_l(b) + \sigma^2_l(w), \quad (1)$$

$$\sigma^2(Y) = \frac{M_h^2}{M^2}\left\{\frac{\sigma_h^2(b)}{n_h} + \frac{\sigma_h^2(w)}{n_h m_h}\right\} + \frac{M_l^2}{M^2}\left\{\frac{\sigma_l^2(b)}{n_l} + \frac{\sigma_l^2(w)}{n_l m_l}\right\} \quad (2),$$

where

$\sigma^2(Y)$	=	Variance for characteristic Y
$\sigma^2_h(b)$	=	Between site variance for high-flow truck stops
$\sigma^2_l(b)$	=	Between site variance for low-flow truck stops
$\sigma^2_h(w)$	=	Within site variance for high-flow truck stops
$\sigma^2_l(w)$	=	Within site variance for low-flow truck stops
M	=	Total number of truck drivers across all truck stops
M_h	=	Number of truck drivers at high-flow truck stops
M_l	=	Number of truck drivers at low-flow truck stops
n_h	=	Number of high-flow truck stops
n_l	=	Number of low-flow truck stops
m_h	=	Number of truck drivers interviewed at high-flow truck stops
m_l	=	Number of truck drivers interviewed at low-flow truck stops

Individual truck stops may be considered to be clusters of truck drivers to which the interview is administered. Formula (2) above does not take into account clustering effects. Such effects tend to decrease variability within clusters, thus increasing sample size as compared to a simple random sample. Clustered samples tend to have less precision than simple random samples of the same size, because units within the same cluster usually are more homogeneous than units from different clusters. Based upon a previous data collection effort at truck stops [Belman 2005], we expect a within-truck stop correlation of responses (ρ) to be less than one percent (0.01). We will make the additional assumptions for sample design optimization purposes that between-site and within-site variances are the same for high-flow and low-flow sites, and that between-site variances are 1/99 times within-site variances (corresponding to a within-site correlation ρ =0 .01). This is summarized as follows:

$$\sigma_h^2(w) = \sigma_l^2(w) = \sigma^2(w) \quad (3)$$

$$\sigma_h^2(b) = \sigma_l^2(b) = \frac{1}{99}\sigma^2(w) \quad (4)$$

The expected variance of the proportion P of truck drivers with given characteristics of interest was calculated according to equation 2 above, accounting for correlation effects in (3) and (4) and costs per survey. Since this study will estimate prevalence of several conditions in the truck driver population, P may vary over a wide range. Consequently, P = 0.5 was used in sample size calculations since that prevalence requires the greatest sample size.

Those values of M, M_h, M_l, m, m_h, m_l, n_h, and n_l corresponding to a standard error ($\sqrt{P(1-P)/n}$)equal to 1.24% were determined. A standard error of 1.24% corresponds to a 95 percent confidence interval for P=0.5 of plus or minus 0.025 (indicating, in this study, a prevalence of a given health condition equal to 50% with 95% confidence interval plus or minus 2.5%). Given the assumptions above (i.e., that the flow of truck drivers in the low-flow sites is about half that of high-flow sites,, that travel costs to low-flow sites are twice those for high-flow sites, and that correlation within truck stops will be 0.01), a total n=2,457 interviews will be needed. Interview administration will take place at 41 high-flow truck stops and 9 low-flow truck stops. Fifty-four interviews are expected to be administered in each high-flow truck stop and 27 administrations at each low-flow truck stop, resulting in a total 2,214 interview administrations at high-flow truck stops and 243 administrations at low-flow truck stops.

Taking into account an expected 20 percent refusal rate for participation in interview administrations [Federal Highway Administration 2002] and estimated 12% driver ineligibility rate [Belman et al 2005], a total of 3,500 truck drivers should take the eligibility screening interview for study participation in order to obtain the 2,457 long-haul truck driver participants needed.

Factors affecting Precision of Estimator

Within-site correlation

In evaluating equation (2) for variance of the estimator, a critical parameter is the within-truck stop correlation coefficient ρ. If the within-site correlation is 0, then estimators from all designs with the same total number of interview administrations will have the same variance. If on the other hand the within-site correlation is high, then designs with the same total number of interviews but taken in a smaller number of sites will show considerably greater variance. The effect of changes in within-site correlation for the design used in this study is shown in Table H1 below. The increase in standard error of the estimator with increasing correlation coefficient ρ is illustrated in Table H1.

Differing Ratios of Drivers Presenting at high- and low-flow truck stops

An important parameter is the ratio of the total number of drivers arriving at high-flow truck stops during the survey period (M_h) to the total number of drivers arriving at low-flow truck stops (M_l). Table H2 presents standard errors using equation (2) for differing ratios of M_h and M_l. The design does well for ratios of M_h to M_l of 4 to 1 or greater, and the standard error is not much higher for somewhat smaller ratios (down to 2 to 1). But if M_l is significantly larger than anticipated (making the ratio significantly smaller and less than 2 to 1), then the variance begins to increase (i.e., the sample size of 9 becomes too small for low-flow sites if the true traffic in these sites is greater than anticipated). A small extra set of low-flow sites will be selected in the National Survey of Truck Driver Injury and Health if it appears that M_l is larger than anticipated.

Additional factors affecting standard error calculations above have to do with extra sources of variability: nonresponse adjustments, variability in the of truck driver counts $\widetilde{M}_{ht}$ and $\widetilde{M}_{lt}$ with time t of the survey period, and T_{ht} and T_{lt} adjustments needed for differences in

field time and staff across sites. Table H2 presents standard errors that include an inflation factor of 50% to account for these extra sources of variability.

Table H1. Expected Standard Error of Estimator for Different Values of Within-Site Correlation ρ

Number of High-Flow Truck Stops	Number of Low-Flow Truck Stops	Number of Interviews at each high-Flow Truck Stop	Number of Interviews at each Low-Flow Truck Stop	Within-Site Correlation Coefficient ρ	Standard Error of Estimator	Relative Increase in Standard Error
41	9	54	27	.01	.0124	
41	9	54	27	.03	.0162	30.5%
41	9	54	27	.05	.0194	56.1%

Table H2. Expected Standard Errors With Varying Percentages of Truck Drivers Stopping at High-Flow Truck Stops

Percent of All Drivers Stopping at High-Flow Truck Stops $M_h/(M_h + M_l)$	Ratio M_h/M_l	Standard error with in-site correlation $\rho=0.01$[1]
98%	49:1	1.59%
95%	19:1	1.55%
90%	9:1	1.52%
85%	5.7:1	1.53%
80%	4:1	1.57%
75%	3:1	1.64%
70%	2.3:1	1.75%
65%	1.9:1	1.87%
60%	1.5:1	2.02%
55%	1.2:1	2.18%
50%	1:1	2.36%
45%	.9:1	2.54%

[1] Includes 50% inflation factor

Sample Selection

A two stage sample selection procedure will be used. The first stage will involve selection of truck stops at which survey administration will occur. The second stage will involve selection of individual truck drivers to whom interviews will be administered.

Truck Stop Selection

The National Survey of U.S. Long-Haul Truck Driver Injury and Health will be conducted at selected truck stops located throughout the 48 contiguous states. These truck stops will be stratified by volume of traffic (high-flow or low-flow) and size of truck stop (number of paved parking spaces for trucks).

High-flow truck stops will be those stops along routes in the National Highway System (NHS) which had (in 2002) a traffic volume of 10,000 trucks or more per day for more than half of their length estimated from the Freight Analysis Framework. The National Highway System includes the Interstate Highway System as well as other roads important to the nation's economy, defense, and mobility, such as:

1. Principal arterials which provide access between an arterial and a major port, airport, public transportation facility, or other intermodal transportation facility.
2. The Strategic Highway Network of highways which provide defense access, continuity and emergency capabilities for defense purposes.
3. Major Strategic Highway Network Connectors
4. Intermodal Connectors

National Highway System routes considered for the National Survey of Truck Driver Injury and Health are listed in Table H3 and shown in Figure H1. The high-flow truck stop sampling frame will be restricted to truck stops along these designated major arteries, which will then be divided into state-artery sections. These state-artery sections will be sampled with probability proportional to the length of the state-artery section. The sampling of state-artery sections will be stratified by region (West, Central, Great Lakes, South, Northeast). Table H4 lists geographic regions included in this study.

Low-flow truck stops will include all truck stops which are not located on the high-flow routes listed in Table H3. They will be selected from truck stops in individual states. Individual states will be sampled with probability proportional to their population. Selecting whole states will avoid needing to designate a set of arteries with truck volume less than 10,000 trucks per day. After excluding the high-flow arteries, it is assumed that other truck traffic in the state will be roughly proportional to population. Low-flow truck stops will be chosen from truck stops in selected states which are not on the high-flow routes listed in Table H3.

Table H3. High Flow National Routes

CA State 99	Sacramento, CA to I-5 Jct south of Bakersfield,
NJ Turnpike	All
I-10	Santa Monica, CA to I-20 Jct Texas
I-10	Jacksonville, FL to San Antonio, TX
I-12	All
I-16	All
I-20	I-10 Jct TX to Shreveport, LA
I-20	Meridian, MS to Birmingham, AL
I-24	All
I-25	Fort Collins, CO to Pueblo, CT
I-26	All
I-30	All
I-35	Wichita, KS to San Antonio, TX
I-40	I-15 jct to I-95 jct
I-44	All
I-45	Dallas, TX to Houston, TX
I-5	All
I-55	Memphis, TN to Chicago, IL
I-57	All
I-65	Nashville, TN to Chicago, IL
I-65	Decator, AL to Mobile, AL
I-69	Indianapolis, IN to Flint, MI
I-70	Baltimore, MD to Grand Junction, CO
I-71	All
I-74	Indianapolis, IN to Cincinnati, OH
I-75	Bay City, MI to Naples, FL
I-76	All
I-77	Charleston, WV to Charlotte, NC
I-80	New York City, NY to Salt Lake City, UT
I-80	Sacramento, CA to Oakland, CA
I-81	Scranton, PA to I-40 junction
I-84	All
I-85	Atlanta, GA to Durham, NC
I-87	Albany, NY to New York, NY
I-90	Albany, NY to Rochester, MN
I-91	New Haven, CT to Hartford, CT
I-94	Minneapolis, MN to Detroit, MI
I-95	Miami, FL to Bangor, ME
I-205	San Lorenzo, CA to I-5 Jct

Table H4. States Included by Geographic Region

Region	States Included
Northeast	Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, Pennsylvania, New Jersey, Delaware, Maryland
Great Lakes	Ohio, Indiana, Michigan, Illinois, Wisconsin, Minnesota
South	Virginia, West Virginia, North Carolina, South Carolina, Georgia, Florida, Alabama, Tennessee, Kentucky, Louisiana, Mississippi, Arkansas
Central	North Dakota, South Dakota, Iowa, Nebraska, Missouri Oklahoma, Kansas, Texas
West	Washington, Idaho, Montana, Wyoming, Oregon, California, Nevada, Utah, Colorado, New Mexico, Arizona

The sample frame for truck stops will be based on listings included in the publication *Trucker's Friend*: *2008 National Truck Stop Directory* [Brice 2008]. Any given truck stop will either be high-flow or low-flow depending on whether it is located on a high-flow national route or not. No truck stop will have two chances of selection.

Truck stops are defined in the *Truckers Friend* by size as follows:

- 'S': 5 to 24 truck parking spaces;
- 'M': 25 to 84 truck parking spaces;
- 'L': 85 to 149 truck parking spaces;
- 'XL': 150 or more truck parking spaces.

One truck stop will be selected with probability proportional to size for each selected state-artery or state. If the same state-artery or state is selected multiple times, as many truck stops as the number of selections of the state-artery or state will be selected; in this case, the selection of truck stops will be without replacement in order to ensure that the same truck stops are not sampled twice. The measures of size defined for each size category will be as follows:

- 'S': measure of size 15;
- 'M': measure of size 55;
- 'L': measure of size 127;
- 'XL': measure of size 250.

If the management of a particular truck stop refuses participation, that truck stop will be replaced with a replacement truck stop within the same artery-state (for the high-flow stratum) or within the same state for the low-flow stratum. The replacement truck stop will have the same measure of size as the refusing truck stop and will be as close as possible to the refusing truck stop (on the same road if possible for the low-flow stratum).

Table H5 illustrates expected numbers of truck stops to be surveyed and numbers of interviews given, assuming 54 interviews at each high-flow truck stop and 27 interviews at each low-flow truck stop.

Table H5. Expected Sample Allocation

Type of Truck Stop	All U.S. Truck Stops[1]	Sample	Number of Interviews
High Flow	1802	41	2214
Low Flow	2412	9	243
Truck Stop Size			
Small	1784	21	1026
Medium	1356	16	783
Large	541	6	297
Extra-Large	533	7	351
Total	4214	50	2457

[1] listings from Trucker's Friend: 2008 National Truck Stop Directory [Brice 2008].

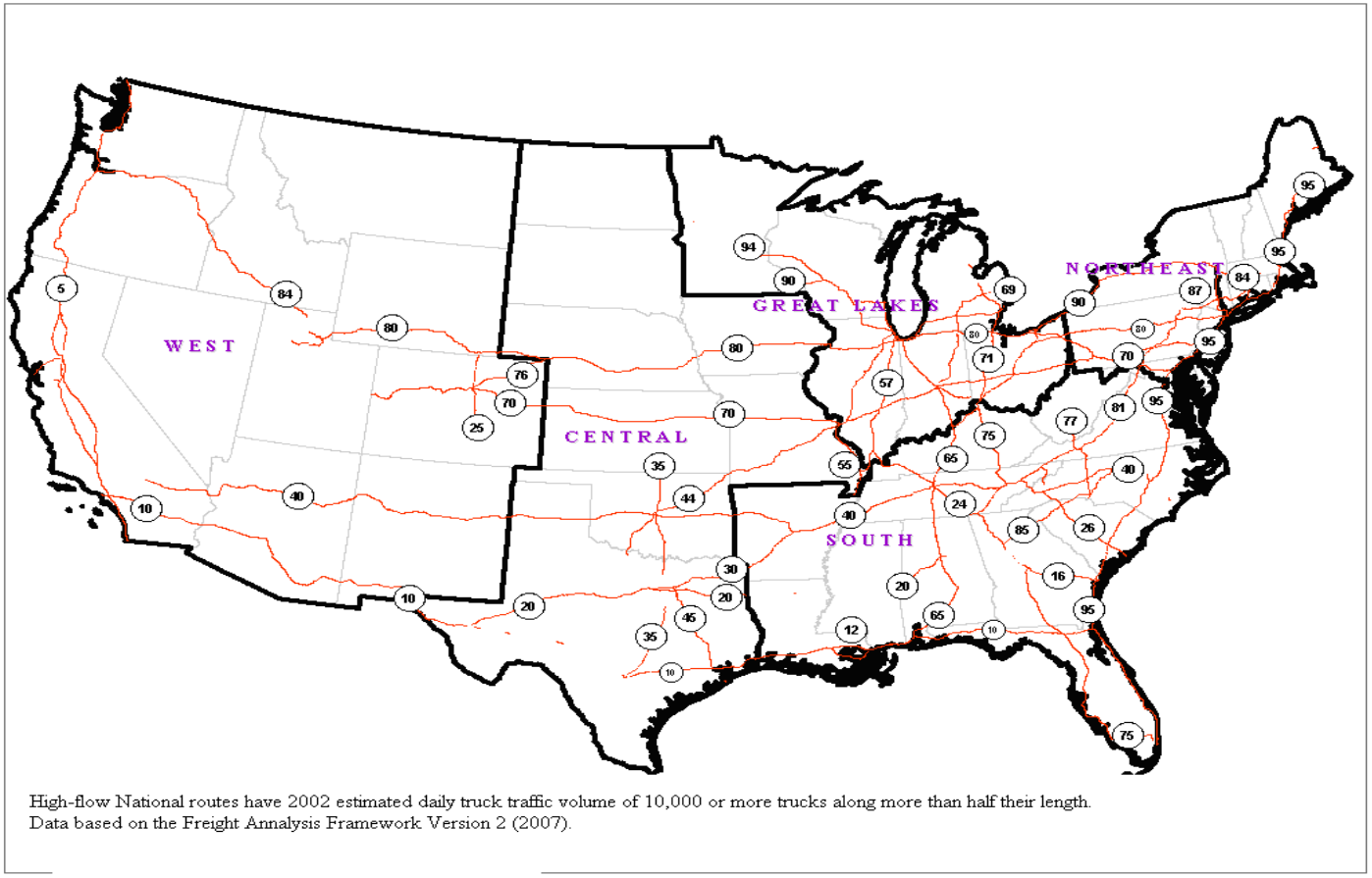

Figure H1. High-Flow National Routes by Geographic Region

Random Selection of Truck Drivers Within Truck Stops

A sample of truck drivers entering the truck stop during periods of data collection will be selected for personal interview. To ensure random selection of drivers, as an interviewer is about to become available the very next driver entering the truck stop will be invited to participate in the study by first taking a screening interview to determine study eligibility. Each truck driver who enters during a period when an interviewer is available for administration will then have an equal chance of selection in this process (i.e., the process of selecting the next driver will be objective and not subject to the discretion of the recruiter). It is also assumed that time entries of drivers into the truck stops are sufficiently inherently random so that there will be no systematic bias generated from not being able to interview drivers who enter when another interview is still being done. There will be no recruitment for interviews during periods when an interviewer is not available. It is anticipated that eligible drivers will have less than a one hour wait before being interviewed. A tally will be kept of all truck drivers entering the truck stop during multiple randomly sampled data collection periods. Recruitment and interview administration periods will include morning and lunch periods (8-10 AM and 12-2 PM) as well as the late afternoon/dinner period.

Data Collection Teams and Training

Survey administration and driver recruitment at each truck stop will be done by a team consisting of three individuals: one member to recruit participants and two members to administer personal interviews and collect anthropometric measurements. Interviews are to be conducted during a 3-day period at each truck stop. During the three day period, truck drivers will be selected randomly to be interviewed during different time periods each day. As the interviewers finish their interviews, the next set of truck drivers will be selected for interviewing.

Three teams of three data collectors will be trained, along with two back-up people for a total of 11 trainees. All surveyors will attend training covering the technical and administrative protocols required to successfully complete the data collection activities, and which will provide them general information about health issues and long haul trucking operations. The training will help data collectors understand the context and purpose of various questions. The following topics will be covered in the training sessions and included in the surveyors' training manual:

- Background on the Survey of Truck Driver Injury and Health;
- Detailed overview of data collection fundamentals;
- Refusal avoidance and conversion techniques;
- Instructions for recording and transmitting data;
- Discussion of Privacy and Human Subjects Rights;

- Overview of health concerns;
- Long haul trucking operations;
- Research focus of the National Institute for Occupational Safety and Health (NIOSH);
- Trucking oversight responsibilities of the Federal Motor Carrier Safety
- Administrative procedures.

Estimation and Weighting

High-Flow Truck Stops

The overall probability of selection p_{hi} for each high-flow state-artery will be

$$p_{hi} = \frac{n_h u_{hi}}{\sum u_{hi}}$$

where u_{hi} is the mileage length of the state-artery section hi and the summation is over all high-flow state-artery sections.

One truck stop will be selected with probability proportionate to size for each selected state-artery segment. The probability of selection p_{hij} of the sampled truck stop is

$$p_{hij} = \frac{s_{hij}}{\sum_{j=1}^{N_{hi}} s_{hij}}$$

where s_{hij} is the measure of size of the truck stop, and N_{hi} is the number of truck stops listed for state-artery unit *hi*.

Low-Flow Truck Stops

The overall probability of selection p_{ls} at the state level for the low-flow sample will be

$$p_{ls} = \frac{n_l Pop_s}{\sum Pop_s}$$

where Pop_s is the 2008 population of the state, and the summation is over all states except Alaska and Hawaii.

Within selected states, one truck stop will be sampled with probability proportional to size (number of parking places). The probability of selection for each low-flow truck stop within each state will be

$$p_{lsj} = \frac{v_{lsj}}{\sum_{j=1}^{N_{ls}} v_{lsj}}$$

where v_{lsj} is the measure of size of the truck stop (15, 55, 127, 250), and N_{ls} is the number of low-flow truck stops listed for state *s*.

Probability of selection of truck drivers

The probability of selection of each truck driver will then be m_{hij}/M_{hij} for high-flow truck stops, where m_{hij} is the number of interviewed truck drivers in the truck stop *hij* and M_{hij} is the estimated number of truck drivers who entered the truck stop *hij* during the interview period. Similarly for low-flow truck stops the probability of selection of each truck driver is m_{lsj}/M_{lsj}, where m_{lsj} is the number of interviewed truck drivers in the sampled truck stop *lsj* and M_{lsj} is the estimated number of truck drivers who entered the truck stop *lsj* during the interview period

Weights

The final weight W_i assigned to each interview will be a product of the following factors:

- A base weight factor equal to the inverse of the probability of selection of the sampled truck stop;
- A base weight factor equal to the inverse of the probability of selection of the truck driver within the sampled truck stop;
- An interviewer-day adjustment factor T (if necessary);
- A nonresponse adjustment NR applied for nonresponding truck drivers;

The nonresponse adjustment will be based on responses to the nonresponse questions and will be the reciprocal of the weighted response rate to selected questions included in the non-respondent interview (**Attachment F2**).

A total of 50 truck stops (41 high-flow and 9 low-flow) will be selected. The mean number of completed interviews expected for high-flow sites will be 54 and the mean number of completed interviews for the low-flow sites is expected to be 27. The actual value for any given truck stop will vary, however. The only value that is fixed under this sample design is the number of days interviews are to be administered at each type of truck stop. If for some reason the number of interviews within the time period varies for

a given site, then an adjustment in the weighting procedure will be made for that site for estimation purposes. For example, suppose the expected number of interviews at a site could not be completed because of the illness of the interviewer during the interview period. If only half of the expected number of interviews were completed, a weighting factor adjustment of 2 would be applied. Likewise, if a given site is visited for two days rather than three for some reason, then a weighting factor adjustment of 1.5 would be attached.

Estimation

If m^{*}_{hij}, m^{*}_{lij} are the respondent sample sizes for sites *hij* (*lij*), the estimators of a particular characteristic *y* may be given as

$$\bar{y}_{hij} = \frac{1}{m^{*}_{hij}} \sum_{k=1}^{m^{*}_{hij}} y_{hijk} NR_{hijk} \qquad \bar{y}_{lij} = \frac{1}{m^{*}_{lij}} \sum_{k=1}^{m^{*}_{lij}} y_{lijk} NR_{lijk}$$

The quantity NR_{hijk} (NR_{lijk}) is the nonresponse adjustment for driver *hijk* (*lijk*). An overall nationally representative estimator for prevalence of a characteristic *y* among long-haul truck drivers across both high- and low-flow truck stops (denoted by subscripts ht or lt) may be given as:

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n_h} \frac{\hat{M}_{hij} \bar{y}_{hij} T_{hij}}{p_{hi} p_{hij}} + \sum_{i=1}^{n_l} \frac{\hat{M}_{lij} \bar{y}_{lij} T_{lij}}{p_{li} p_{lij}}}{\sum_{i=1}^{n_h} \frac{\hat{M}_{hij} T_{hij}}{p_{hi} p_{hij}} + \sum_{i=1}^{n_l} \frac{\hat{M}_{lij} T_{lij}}{p_{li} p_{lij}}} = \frac{\hat{Y}}{\hat{M}} \quad ,$$

where the quantities $\hat{M}_{hij}$ ($\hat{M}_{lij}$) are estimates based on short-period counts of the number of truck drivers who pass through truck stop *hij* (*lij*) during the data collection period, and T_{hij} (T_{lij}) is an adjustment to a common length period for all truck stops (if the length of time varies from the norm of three days).

We will estimate variance by using replication, given the complexity of an exact variance formula. Equation (2) can be seen as an approximation if we leave out extra components of variability contributed by $\hat{M}_{ht}$ ($\hat{M}_{lt}$) and T_{ht} (T_{lt}), as well as the nonresponse adjustments NR_{htk} (NR_{ltk}).