Special Surveys Division
Division des enquêtes spéciales
Ottawa, Ontario, Canada K1A 0T6

THE DRINKING AND DRIVING

MICRO DOCUMENTATION AND USERS GUIDE

September, 1988

Statistics Canada Statistique Canada

Canada

TABLE OF CONTENTS

1. INTRODUCTION

This package was designed to enable interested users to access and manipulate the microdata file for the Survey of Drinking and Driving, conducted in March 1988. It contains information on the objectives, methodology and estimation procedures as well as guidelines for releasing estimates based on the survey.

Appendix A contains the record layout, the major part of this documentation package. The survey questionnaires are contained in Appendix C.

Any questions about the data set or its use should be directed to:

Statistics Canada
Special Surveys Division
5th floor, Section B6
Jean Talon Building
Ottawa, Ontario
1-800-461-9050
fax (613) 951-0562
special@statcan.ca

2. OBJECTIVES

The statement of objectives for the survey as outlined by Health and Welfare Canada were:

1. To collect etiological attitudinal, cognitive and behavioral information regarding drinking and driving.

2. To collect information that is representative (of the population with telephones and non-institutionalized) and useful at both the provincial and national levels.

3. To collect baseline data which can be used to assess trends and changes in variables over time (follow-up survey in several years).

The goal of the overall program (of which the national survey was a part) is to change the prevailing social acceptability of drinking and driving at the community level.

3. POPULATION

The target population for the Drinking and Driving Survey was all persons aged between 16 and 69 inclusively in Canada excluding:

1. residents of the Yukon and Northwest Territories;

2. full-time residents of institutions.

Households were surveyed through random digit dialing (RDD), a telephone sampling method. As a result, households without telephones were excluded. These represent less than 3% of the population.

The survey estimates have been adjusted (weighted) to represent persons without telephones.

4. SURVEY DESIGN

The Drinking and Driving employed two different Random Digit Dialing sampling techniques.

For Newfoundland and Ontario, the Elimination of Non Working Banks method was used. A description of this procedure is given in Section 9.2. For the remaining provinces the Waksberg method was used. This procedure is described in Section 4.1.

4.1 Telephone (Waksberg Design) (P.E.I., N.S., N.B., Que, Man., Sask., Alta., B.C.)

The Waksberg Method is a Random Digit Dialing sampling technique which significantly reduces the cost of a survey as compared to dialing telephone numbers completely at random. The method employs a two-stage design which increases the likelihood of contacting households. The following describes what was done for the Drinking and Driving Survey in the above-mentioned provinces for the 16-69 age group.

First, each of the eight provinces was divided into strata — one stratum representing the Census Metropolitan Area (CMAs) of the province, the other representing the non-CMA areas. Montreal formed a separate stratum. An up-to-date list of all telephone area code and existing prefix number combinations was obtained for each stratum. To these, all possible combinations of the next two digits were added (i.e., all possible banks of 100 consecutive numbers within existing area code - prefix combinations were identified). This resulted in a list of all the possible first eight digits of ten digit telephone numbers in each stratum. These eight digit numbers formed the first stage sampling units (i.e., were the Primary Sampling Units - PSUs).

Within each stratum, a random selection was made of one of these eight digit numbers and then the final two digits were generated at random. This number (called a Primary number) was called to determine whether or not it reached a household. If it did not reach a household (i.e., the number was not assigned for use or was a business, institution, etc.), the number was dropped from further consideration. If it did reach a household, additional numbers (referred to as secondary numbers) were generated within the same bank (i.e., numbers with the same first eight digits as the primary number). These numbers were also called to determine whether or not they reached a household.

Secondary numbers were generated on a continuing basis until (i) five additional households were reached in each retained bank or (ii) the bank was exhausted or (iii) the survey period ended.

Primary numbers were generated continuously throughout the survey period in order to yield a predetermined required number of households within each stratum. An attempt was made to conduct an interview with a randomly selected respondent in all primary and secondary households reached in which there was at least one household member in the 16-69 age group. To select a person in the household, a grid was attached to each questionnaire. The grid had two lines, one representing the household size and one giving the rank on the list of the person that was to be interviewed. The people were always listed the same way, starting with the oldest person of the household. The ranks were randomly generated on each grid.

This method is more efficient than "pure" random digit dialing because there is a higher probability of reaching a household if the telephone number for that household is within a bank

that contains at least one other household. For the Drinking and Driving Survey, in the provinces in which the Waksberg method was used, about 56.1% of the secondary numbers called reached a household, while only 18.5% Of the primary numbers called reached a household.

4.2 Telephone Sample (Elimination of Non-Working Banks Design) Persons 16 - 69 (NFLD. & Ont.)

The Elimination of Non-Working Banks design is a form of Random Digit Dialing in which an attempt is made to identify all working banks for an area i.e., to identify all banks that have within them at least one household. Thus, all telephone numbers within non-working banks are eliminated from the sampling frame. This is the method that was used to sample the 16-69 age group in Newfoundland and Ontario. Again, strata were formed within each of the two provinces representing CMA and non-CMA portions. Toronto formed a separate stratum.

In Newfoundland, there were two telephone companies. They provided to Statistics Canada a typed list of banks, that was then keyed in to the computer system. The first list had 1452 banks and he second one, 2373. 60 banks were removed from the second list because they were known to be business only all possible two last digits were randomly generated for each bank to create the frame.

A systematic sample of telephone numbers was then generated for each stratum within these working banks. The entire sample of telephone numbers was generated on the first day of interviewing. Therefore, a prediction had to be made of the percentage of numbers dialed that would reach a household. This is what is known as the "hit rate". In addition to this a prediction had to be made of the percentage of households that would contain at least one person in the 16-69 age group. Approximately 2707 telephone numbers were generated or the sample with the expectation that this would result in around 1236 households being contacted 1000 of which would be eligible responding households. This represents a hit rate of approximately 45.2%.

In actuality, a hit rate of 46.3% was achieved; 1333 households were contacted of which 1002 were eligible and responding. The hit rate was predicted from a previous Random Digit Dialing Survey which employed the ENWB method.

As was done for the other eight provinces, an attempt was made to conduct a DAD interview with one randomly selected person between the ages of 16 and 69 from each eligible household reached.

Because Bell Canada does not service all of Ontario, the list of area code-prefix-banks obtained from the Bell Canada files was supplemented with the list of all possible area code-prefix-banks combination in non-Bell areas. This procedure has the effect of somewhat lowering the expected hit rate, but because the non-Bell areas represent a small portion of Ontario population, the resulting hit rate is still quite acceptable.

As was done for Newfoundland, a systematic sample of telephone numbers was then generated from the telephone numbers within these working banks. Approximately 2403 telephone numbers were generated with the expectation that this would result in around 1289 households being contacted, 1000 of which would be eligible responding households. This represents a hit rate of around 50.3%. In actually, a hit rate of 53.6% was attained; 1306 households were contacted, 1010 of which were eligible and responded.

5. COLLECTION

Data collection for the Drinking and Driving Survey was conducted by telephone. There was one control form to record some household composition information (SDD 01) and a questionnaire that was administered to a person chosen at random from the eligible household members. Questionnaires and procedures were field tested in a pretest involving approximately 375 households in July 1987.

Telephone data collection took place during the period February 29 to March 19, 1988. All telephone interviewing took place from centralized telephone facilities in Statistics Canada's regional offices with calls being made from approximately 9 a.m. until 9:30 p.m. Interviewers were trained by Statistics Canada staff in telephone interviewing techniques, survey concepts and procedures during an 8 hour classroom training session. The majority of interviewers had previous telephone interviewing experience. Interviewer supervisors were responsible overseeing the quality of interviewer's work.

Listed below is a list of the manuals used in the survey. Copies of these are available on request.

SDD03 Survey on Drinking and Driving Interviews Manual

SDD06 Survey on Drinking and Driving Procedures Manual

SDD08 Survey on Drinking and Driving Training Guide

6. PROCESSING

The following is an overview of the processing steps for the survey.

6.1 Data Capture

Data from the survey questionnaires were entered directly into mini-computers in Statistics Canada's regional offices and transmitted to Ottawa. The data capture program allowed for a valid range of codes for each question and automatically followed the flow of the questionnaire.

6.2 Edit and Imputation

All survey records were subjected to an exhaustive computer edit to identify and correct invalid or inconsistent information on the questionnaires. Records with missing or incorrect information were assigned non-response codes or corrected from other information from the respondents' questionnaire.

6.3 Imputation

Due to the nature of the survey, imputation was not appropriate for most items and thus "not stated" codes were usually assigned for missing data. However, non-response was not permitted for those items required for weighting. For number: of telephones in the household, a value of 1 was assigned automatically the variables age, sex and household, size were imputed randomly.

The table below indicates the number of records requiring imputation for each item on the survey.

Variable	Records requiring Imputation
Age	72
Sex	85
household size	19

7. ESTIMATION

The principle behind the estimation procedure when a probability sample is used, as was used for the Drinking and Driving Survey, is that each person selected in the sample "represents", besides himself/herself, several other persons not in the sample. For example, in a simple random sample of 2% of the population, each person in the sample represents 50 persons in the population.

For the microdata file that was created for the Drinking and Driving Survey, there is one record for each person who responded to the survey. Each record contains demographic characteristics of the selected person as well as the questionnaire information on drinking and driving. Instead of physically duplicating the sample records according to the number of persons the records represent, an overall weighting factor was placed on each record. The weighting factor refers to the number of times a particular record should be replicated to obtain populations estimates. For example, if the number of persons who have not had a drink in the last year is to be estimated, this is done by selecting the records on the microdata file who reported that they did not drink in the last twelve months and summing the weights on these records.

7.1 Weighting

Because the Drinking and Driving Survey employed sampling techniques (as discussed in Section 4), two slightly different weighting procedures were employed. Each briefly in the following sections.

7.1.1 Waksberg Design (16-69 P.E.I., N.S., N.B., Que., Man., Sask., Alb., B.C.)

A self-weighting sample design is one for which the weights for each unit in the sample are the same. For a two-stage sample design, this happens proportional to size (PPS) sampling and a fixed number of units are selected within each selected Primary Sampling Unit with equal probability.

For the Drinking and Driving Survey, the 16-69 age group households within the strata in each of the eight above-mentioned provinces were selected using this sampling scheme and thus, the sampled households within each stratum have identical weights. The first stage sampling units (i.e., the Primary Sampling Units) were banks of telephone numbers and the second stage units were actual telephone numbers corresponding to households within those banks. It should be noted that household weights differ from province to province because a different sampling rate was used for each province.

The following outlines the steps that were used in weighting the DAD Waksberg records for the 8 provinces.

1) Basic Weight

In the first stage of weighting all households that were selected into the sample within a given stratum (within a province) were assigned an identical weight of 1.

2) Non-Response Adjustment

Weights for responding households were adjusted to represent non-responding households. Within each working bank of telephone numbers selected in the sample, the sampling scheme used required that six households be contacted. In some cases one or more of these six households refused to participate in the survey. Weights of responding households were adjusted to compensate for non-responding households by multiplying the basic weight of responding households within a bank by the following ratio:

$$\frac{6}{\text{No. of responding households within the bank}}$$

Note: There have been a few banks where because of constraints, less than 6 numbers were generated. The 6 was decreased accordingly.

3) Multiple Telephone Adjustment

Weights for households with more than one private telephone number were adjusted downwards to account for the fact that such households have a higher probability of being selected. The weight for each household was divided by the number of distinct telephone numbers that serviced the household (as determined in the interview).

4) Person Weight Calculation

A person weight was then calculated for each person who responded to the survey by multiplying the household weight for that person by the number of persons in the household who were eligible to be selected for the survey (i.e., the number of household members between the age of 16 and 69).

5) Adjustment for External Totals

An adjustment was made to the person weights on records within each stratum in order to make population estimates consistent with Census projected population counts. This was done by multiplying the person weight for each record within the stratum by the following ratio:

$$\frac{\text{Census population project for the stratum}}{\text{sum of the person weights of D \& D Waksberg records within the stratum}}$$

6) Province - Age - Sex Adjustment

Again the final weighting step was to ratio adjust the weights of Waksberg records to agree with Census projected population counts were obtained for males and females within the following age groups: 6-19, 20-24, 25-34, 35-44, 45-54, 55-64, 65-69. For each of tine resulting 112 classifications (8 X 7 X 2) the persons weights for records within the classification were adjusted by multiplying by the following ratio:

$$\frac{\text{projected census population count}}{\text{sum of the person weights of records in the prov - age - sex group}}$$

It should be noted that persons living in households without telephone service are included in these projections even though such persons were not sampled.

7.1.2 Elimination of Non Working Banks Design (16-69) Newfoundland and Ontario

As was the case for the Waksberg design, when the Elimination of Non-Working Banks (ENWB) design is used, each household within a stratum has an equal probability of selection. This probability is equal to:

$$\frac{\text{No. of telephone numbers sampled within the stratum}}{\text{total number of possible telephone numbers within the stratum}}$$

(The total number of possible telephone numbers for a stratum is equal to the number of working banks for a stratum times 100). The following steps outline the weighting procedure that was used for ENWB records.

1) Basic Weight

Each household (responding and non responding) was assigned a weight equal to the inverse of its probability of selection:

$$\frac{\text{total number of possible telephone numbers within the stratum}}{\text{No. of telephone numbers sampled within the stratum}}$$

TABLE 1 RESPONSE RATES

Province	Hit Rate (overall)	number of non-households	number of 0.0 scope households	number of ring / no answer	ring / no answer rate	number of non-response refusal / not at home	non-response rate	response	response rate
NFLD.	46.3%	1549	104	123	10.0%	104	8.5%	1002	81.5%
P.E.I.	35.2%	2368	158	43	3.8%	82	7.3%	1001	88.9%
N.S.	40.8%	1866	138	54	4.7%	97	8.3%	999	87.0%
N.B.	40.1%	1871	123	57	5.0%	105	9.3%	968	85.7%
P.Q.	45.6%	1511	103	59	5.1%	117	10.0%	989	84.9%
ONT.	53.6%	1132	145	70	6.0%	81	7.0%	1010	87.0%
MAN.	30.4%	2912	158	18	1.6%	145	13.0%	952	85.4%
SASK.	27.7%	3523	180	30	2.6%	104	8.9%	1035	88.5%
ALB.	39.3%	1941	72	33	2.8%	151	12.7%	1002	84.5%
B.C.	49.7%	1328	123	45	3.8%	161	13.5%	985	82.7%
TOTAL	39.3%	20001	1304	532	4.6%	1147	9.8%	9943	85.6%

$$\text{Note: hit rate} = \frac{\#\,\text{hhlds(in scope} + 0.0\,\text{scope)}}{\#\,\text{hhlds(in scope} + 0.0\,\text{scope)} + \#\,\text{not households}}$$

2) Non-Response Adjustment

Weights for responding households were adjusted to represent non-responding households. This was done independently within each area code prefix. Records were adjusted by the following factor:

$$\frac{\text{sum of the household weights of all households within the area code}}{\text{sum of the household weights of responding households with the area code prefix}}$$

Non-responding, households were then dropped. The next 4 steps used in the weighting procedure for ENWB records were identical to what was done for Waksberg records (i.e., see Waksberg Section 8.2.2 for details).

3) Multiple Telephone Adjustment

4) Person Weight Calculation

5) Adjustment to External Totals

6) Province - Age - Sex Adjustment

7.1.3 Section on Non-response rates

Different rates are provided in Table 1. The overall hit rate of the survey was 39.3% (as mentioned earlier, the hit rate represents the proportion of households reached from all the dialed telephone numbers). It can be seen that hit rates are higher for the provinces that used the ENWB technique than for those using Waksberg technique. The second column gives the number of calls that reached non-households (ex: institutions, businesses, numbers not in service...). The columns 3 to 9 are giving information about the households. The third column represents the number of telephone calls that reached households out-of-scope for this survey. These are mostly households with persons aged 70 years and more only. The last six columns represents response/non-response counts and rates. The non-response/refusal/not at home category includes both the household non-responses and the selected person non-responses. Reasons for non-responses are refusal, illness, injury, absence for the duration of the survey, language problem....

7.2 Weighting Policy

Users are cautioned against releasing unweighted tables or performing any analysis based on unweighted survey results. As was discussed in Section 7.1, there were several weight adjustments performed independently on records within each province. Sampling rates as well as non-response rates varied significantly from province to province.

It is known that non-respondents are more likely to be males and more likely to be younger (20 - 24). In the responding sample, 4.01% were males between the ages of 20 and 24, while in the overall population, approximately 6.58% are males between 20 and 24. Therefore, it is clear that the sample counts cannot be considered to be representative of the survey target population unless appropriate weights are applied.

7.3 Types of Estimates

Two types of "simple" estimates are possible from the results of the Drinking and Driving Survey. These are qualitative estimates (estimates of counts or proportions of people possessing certain characteristics) and quantitative estimates involving quantities or averages. More complex estimation and analyses are covered in section 8.4.

7.3.1 Qualitative Estimates

It should be kept in mind that the target population for the DAD was non-institutionalized persons aged 16 - 69 living in the ten provinces. Qualitative estimates are estimates of the number or proportion of this target population possessing certain characteristics. The number of women living in Ontario who drink is an example of this kind of estimate. These estimates are readily obtained by summing the final weights of the records possessing the characteristic in questions.

7.3.2 Quantitative Estimates

Some variables on the Drinking and Driving Survey microdata file are quantitative in nature (e.g., number of drinks in last seven days, weight and age). From these variables, it is possible to obtain such estimates as the average number of drinks in the last seven days of people who drive more than 20,000 km per year. These estimates are of the following ratio form:

$$\text{est(average)} = \frac{X}{Y}$$

The numerator (X) is a quantitative estimate of the total of the variable of interest (average number of drinks in last seven days). The denominator (Y) is the qualitative estimate of the number of participants (who drive more than 20,000 km per year). The two estimates are derived independently and then divided. For the example given, X would be calculated by multiplying the final weights on records (referring to persons) reporting they drove more than 20,000 km/year by the value given for average number of drinks and summing these products over all applicable records. The value Y would be calculated simply by summing the final weights of records reporting they drive more than 20,000 km per year.

7.4 Guidelines for Analysis

As is detailed in section 4 of this document, the respondents from the DAD do not form a simple random sample of the target population. Instead, the survey had a complex design, with stratification and multiple stages of selection, and unequal probabilities of selection of respondents. Using data from such complex surveys presents problems to analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used.

The DAD used a stratified design, with significant differences in sampling fractions between strata. Thus some areas are over-represented in the sample (relative to their populations) while some other areas are relatively under-represented; this means that the unweighted sample is not representative of the target population. The survey weights must be used when producing estimates or performing analyses in order to account for this over- and under-representation. While many analysis procedures found in statistical packages allow weights to be used, the

meaning or definition of the weight in these procedures differ from that which is appropriate in a sample survey framework, with the result that while in many cases the estimates produced by the packages are correct, the variances that are calculated are almost meaningless.

For many analysis techniques (for example linear regression, logistic regression, estimation of rates and proportions and analysis of variance) a method exists which can make the variances calculated by the standard packages more meaningful. If the weights on the data are rescaled so that the average weight is one (1), then the variances produced by the standard packages will be more reasonable; they still will not take into account the stratification and clustering of the sample's design, but they will take into account the unequal probabilities of selection. The rescaling can be accomplished by dividing each weight by the overall average weight before the analysis is conducted.

The calculation of truly meaningful variance estimates requires detailed knowledge of the design of the survey; such detail cannot be given in this microdata file because of confidentiality. Variances this take into account the sample design can be calculated for many statistics by Statistics Canada on a cost recovery basis.

8. RELEASE GUIDELINES AND DATA RELIABILITY

It is important for users to become familiar with the contents of this section before publishing or otherwise releasing any estimates derived from the DAD Survey microdata file.

This section of the documentation provides guidelines to be followed by users. With the aid of these guidelines, users of the microdata should be able to produce figures consistent with those produced by Statistics Canada and in conformance with the established guidelines for rounding and release. The guidelines can be broken into two broad sections - sampling variability and rounding policy.

8.1 Sampling Variability Guidelines

The estimates derived from this survey are based on a sample of households. Somewhat different figures might have been obtained if a complete census had been taken using the same questionnaire, interviewers, supervisors, processing methods, etc. as those actually used. The difference between the estimates obtained from the sample and the results from a complete count taken under similar conditions is called the sampling error of the estimate.

Although the exact sampling error of the estimate, as defined above, cannot be measured from sample results alone, it is possible to estimate a statistical measure of sampling error, the standard error, from the sample data. Using the standard error, confidence intervals for estimates (ignoring the effects of nonsampling error) may be obtained under the assumption that the estimates are normally distributed about the true population value. The chances are about 68 out of 100 that the difference between a sample estimate and the true population value would be less than one standard error, about 95 out of 100 that the difference would be less that two standard errors, and virtually with certainty that the differences would be less then three standard errors.

Because of the large variety of estimates that can be produced from a survey the standard deviation is usually expressed relative to the estimate to which it pertains. The resulting measure, known as the coefficient of variation or an estimate is obtained by dividing the standard error of the estimate by the estimate itself and is expressed as a percentage of the estimate. Before releasing and/or publishing any estimates from the microdata file, users should determine whether the estimate is releasable based on the following guidelines:

8.2 Estimates of Variance

Variance estimation is described separately for qualitative and quantitative estimates.

8.2.1 Sampling Variability for Qualitative Estimates

Derivation of sampling variabilities for each of the estimates which could be generated from the Drinking and Driving Survey would be an extremely costly procedure, and for most users, an unnecessary one. Consequently, crude measures of sampling variability, in the form of tables, have been developed for use and are included in Appendix C (Crude Sampling Variability Tables). These tables have been produced using the coefficient of variation formula based on a simple random sample. Because estimates for the Drinking and Driving Survey were not, in fact, based, on a simple random sample design and were, instead, based on two different sample designs (the Elimination of Non Working Banks design and the Waksberg design), a factor called

the design effect was introduced into the variance formula. The design effect for an estimate is the actual variance for the estimate (taking into account the design that was used) divided by the variance that would result if the estimate had been derived from a simple random sample. For each province and Canada, average design effects were calculated taking into consideration several of the estimates produced by the survey. In appendix B, the Crude Sampling Variability Tables for the 16 - 69 age group were based on these average design effects

TYPE OF ESTIMATE	**COEFFICIENT OF VARIATION (IN %)**	**GUIDELINE STATEMENT**
1 Unqualified	0.0 to 16.5%	Estimates can be considered for general unrestricted releases. No special notation is required, although the alphabetic indicators at left are suggested.
2 Qualified	16.6 to 25.0%	Estimates can be considered for general unrestricted release should be accompanied by warning of high sampling variability associated with the estimates. Such estimates should be identified by the letter G (or some other similar as fashion).
3 Restricted	25.1 to 33.3%	Estimates can be considered for general unrestricted release only when sampling variabilities are obtained using an exact variance calculation procedure. The estimates should be accompanied by a warning of high sampling variability associated with the estimates.
4 Not for Release	33.4% or over	Estimates should not be released in any form under any circumstances. In such statistical tables, such estimates should be deleted.

8.3 Rounding

In order that estimates produced from the Drinking and Driving Survey microdata file correspond to those produced by Statistics Canada users are urged to adhere to the following guidelines regarding the rounding of such estimates. It is unwise to release unrounded estimates, as they imply greater precision than actually exists.

8.3.1 Rounding Guidelines

(1) Estimates of totals in the main body of a statistical table should be rounded to the nearest thousand using the normal rounding technique (see definition in section 8.3.2).

2) Marginal sub-totals and total in statistical tables are to be derived from their corresponding unrounded components and then are to be rounded themselves to

the nearest thousand units using normal rounding.

3) Average, proportions, rates and percentages are to be computed from unrounded components and then are to be rounded themselves to one decimal using, normal rounding.

4) Same and differences or aggregates and ratios are to be derived from corresponding unrounded components and then rounded to the nearest thousand units or the nearest one decimal using normal rounding.

5) In instances in which, due to technical or other limitations, a different rounding technique is used, which results in estimates being released which differ from the corresponding estimates produced by Statistics Canada, users are encouraged to note the reason for such differences in the release document.

8.3.2 Normal Rounding

In normal rounding, if the first or only digit to be dropped is 0 to 4, then last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is raised by one. For example, the number 8499 rounded to thousands would be 8 and the number 8500 rounded to thousands would be 9.