A MULTI-STAGE PROBABILITY SAMPLE FOR TRAFFIC SURVEYS*

Leslie Kish, Survey Research Center, University of Michigan
Warren Lovejoy and Paul Rackow, The Port of New York Authority

Introduction

Origin-destination surveys of vehicular traffic serve to obtain data on the patterns of vehicular travel for use in planning the location, size configuration and competitive effect of future vehicular facilities. Since 1930, The Port of New York Authority has conducted such surveys periodically at the tolls plazas of its tunnels and bridges across the Hudson River on a "one shot" basis, i.e., as many interviews as possible were obtained over a period of one to three days and the results were assumed to be representative of the entire year's traffic.

Because of the deficiencies in this method, a study of the problems and the feasibility of using a probability sample was followed in January 1958 by the adoption of a new sampling technique, dubbed "Continuous Sampling." Briefly, this technique is based on a carefully designed and controlled probability sample which builds up enough interviews over time to give sufficient reliability for the desired estimates.

Estimating the characteristics of vehicular traffic on the six facilities poses some interesting problems in survey sampling. The traffic volume varies greatly among the different facilities as well as over different time intervals. This dimension of time has several identifiable sources of variation: seasonal over months, the pattern over days of the week and the diurnal fluctuation. But we wanted to avoid reflecting these extreme variations in the sampling probabilities for vehicles because the survey objectives made it advantageous to aim for equal probabilities.

Eight hour periods over a single facility, for which estimates of the variable traffic volume are available, served as primary sampling units. Despite these variable volumes we designed roughly equal sampling workloads to make the demands on the enumerator both reasonable and efficient. Equal over-all selection probabilities and equal sampling loads can both be achieved by first selecting the primary sampling units with probabilities proportional to estimated traffic and then sampling within those units with rates inversely proportional to the same probability.

Many other traffic problems possess similar essential features. For example, the traffic over roads crossing the boundaries of a state (or other defined area) has similar dimensions of variations over space and of the various types of variation over time. In still other applications, instead of space and time, other dimensions can be treated similarly. Hence, a description of the solution of such a problem may be of general interest.

Advantages of the Continuous Probability Sample

A number of important gains are achieved by this new traffic survey technique.

(1) By virtue of the fact that the laws of probability hold for this type of sample, we could set confidence limits on the estimates made from the results. In contrast the judgment samples used before could not yield valid inferences about the population in the statistical sense.

(2) Spreading the sampling over time avoids the possibility inherent in "one shot" surveys that seasonal or other variations might make traffic patterns on the day or days surveyed quite unrepresentative of the total population.

(3) This new technique provides up-to-date information on origin and destination patterns at all times because the survey data is continuously processed, tabulated and analyzed. In particular, this method of sampling provides measurements of seasonal variation which has been unavailable heretofore.

(4) The design of the survey eliminates the necessity of hiring large numbers of unskilled, temporary employees as interviewers and coders. For the current survey, four regularly employed interviewers did all of the interviewing as well as the coding. These men were trained in interviewing methods, had a thorough knowledge of the geography of the New York Metropolitan area, and were skilled in the coding of the data.

(5) Actual over-all costs for a year's survey using a probability sample is less than that for the judgment survey taken on three days in the past.

(6) Intensive sampling at any facility at any one time is avoided, thereby reducing the instances of congestion caused by obtaining origin and destination information. Furthermore, skilled interviewers reduce the time required for each interview, reduce the proportion of nonresponse, and obtain more complete information on each interview without undue delay to the motorist.

Data Obtained

During the year 1958, some 92,300 interviews were obtained at the six Port Authority vehicular facilities representing about 0.1 per cent of total annual traffic involved in the survey. A response rate of nearly 99 per cent was achieved which was a considerable improvement over the level obtained in the "one shot" surveys conducted in the past.

*Presented at the 1960 Annual meeting of the American Statistical Association, but did not appear in the 1960 Proceedings.

While the basic purpose of the survey was to obtain information as to the origins and destinations of autos, trucks, and tractor-trailers using the Port Authority tunnels and bridges, supplementary information was obtained concerning purpose of trip, license plate of vehicle, number of passengers in vehicles and time, day-type and direction. Buses were not included in the survey.

The survey was continued through 1959 and is in current use in 1960 with some modifications in the sample design to provide greater statistical reliability for certain types of peak period data.

Survey Requirements

The sample design had to be both economical and practical in order to meet the necessary restrictions inherent in this type of field survey especially as regards reasonable workloads, conditions of work and cost.

There were a number of specific requirements which the survey design had to accomplish:

(1) Sample estimates had to be produced for a great variety of specified lines of travel of varying volumes. A line of travel is defined as traffic moving between two geographically specified zones. These estimates had to be provided not only for the six Port Authority vehicular crossings combined, but for each facility separately on an annual basis as well as for each of the four seasons of the year.

(2) The sampling frame included all revenue autos, trucks and tractor-trailers using the six Port Authority vehicular crossings on all types of days (weekdays, Saturdays, Sundays and holidays). Buses were not included in the frame.

(3) The sample was to be a probability sample requiring that every vehicle trip made during the year have a known probability of selection, achieved by randomized selection procedures at every stage of the sample design.

(4) There had to be built into the sample design a replicative procedure so that proper measurements of the standard errors of the various sample estimates could be made.

(5) The statistical reliability required in the survey specified that with the expected over-all annual sample of 100,000 interviews, a sample estimate of 1 per cent, or 1,000 interviews, would have a coefficient of variation of the order of 3 to 6 per cent.

The design also had to be fitted to the practical problems inherent in the actual conduct of this type of survey, and these considerations necessitated the use of the following procedures:

(6) The survey budget permitted the assignment of four permanent field interviewers to this project. These interviewers were to be responsible for both the collection and the coding of the data obtained in the field. This limitation of available manpower restricted the number of tours of duty in the field to eleven per week.

(7) Reasonable working hours and the desirability of reducing variations in the hourly loads within a shift were used as criteria for designating the tours of duty at a facility. Thus, the shift, an eight hour tour of duty at one facility on one of the seven days of the week, became the primary sampling unit. The tours were designated as, a) 11 p.m. - 7 a.m. b) 7 a.m. - 3 p.m. c) 3 p.m. - 11 p.m.

(8) To improve coverage or different origin and destination patterns, directional flows, vehicle types and vehicular volumes at each facility during an eight hour period, the interviewer moves from one traffic lane to another each hour in a prescribed pattern of rotation. The design includes specified relief periods each hour and it is assumed, with good reason, that the traffic patterns during these relief periods are not different from those sampled during the interviewing periods.

(9) Because of the distances involved in covering the entire roll plaza of some facilities, it was necessary to subdivide some facilities into two or more locations. The locations were selected in such a manner as to allow interviewers to count the number of toll lanes open within a location during each hour of interviewing. The interviewer rotates among the open lanes at one prespecified location for four hours and then moves on to another location.

(10) Experience with previous "one shot" surveys supplemented by field testing indicated that an average workload of forty interviews per hour could be achieved by an interviewer. This is an average rate that can be maintained for long periods and still permitting considerably greater rates for occasional peak hours.

(11) Available data on hourly traffic volumes via each facility on the different days of the week were utilized in selecting uniform sampling rates during each shift. The probability of selecting a shift at a facility was made roughly proportionate to the average hourly density of traffic expected during that shift. Then the selection rates within the shifts were made inversely proportional to that same probability, thus producing equal final probabilities and roughly equal hourly workloads.

(12) Flexibility is required to handle such

factors as the ratio of actual to expected lanes open, loss of interviewing time due to bad weather, unusual variations in volume and nonresponse. The self-weighting sample was maintained by introducing a procedure of balancing actual against expected interviews by duplicating or omitting some as required.

Four-Stage Probability Sample

The sampling plan of the survey follows a stratified multi-stage probability design. The sample was selected in four stages with the probability of selection of any vehicle crossing a facility as the product of the probabilities in the several stages. These probabilities were controlled to maintain the equal chance of selecting all vehicles at any facility at any time. The four stages of selection were:

(1) Shift - an eight-hour tour of duty at one of the four facilities on a given day of the week.

(2) Location - a geographically contiguous group of toll lanes at a facility at which the interviewer remains for four hours.

(3) Lane - one of the toll lanes within the location.

(4) Vehicle - an auto, truck, or tractor-trailer passing through a lane.

The primary sampling units (shifts as defined above) were grouped into 84 strata consisting of four facilities by seven day types by three tours of duty. Each stratum was represented by the measure of size fj, a probability proportional to the expected hourly density of traffic flow for that stratum. Each fj, the approximate thousands of vehicles per hour, was an integer for ease of selection and balance. The sum of the fj's for all strata was 286 (26 weeks x 11 shifts) for each half year period, divided into 143 for two seasons. A table of selection was made up of the fj for each of the 84 strata and systematic random sampling was used to make up 13 weekly workloads of 11 shifts each with each workload roughly spread across the strata. Then the workloads were assigned at random to the 13 weeks in the season.

Probabilities for the Four Stages of Selection

The probability of selecting a vehicle crossing a facility may be represented as:

$$\Pr(\text{Selecting a Vehicle}) = \left(\frac{11\ fj}{286}\right)\left(\frac{1}{24\ fj}\right)\left(\frac{3}{4}\right) = \frac{1}{832}$$

The first term on the right denotes the probability of selecting a shift. This is the first stage of selection. Eleven shifts are selected for any week with probability fj/286.

The third term, the fraction 3/4, reflects the fact that interviewing takes place only 44 minutes out of each hour $\left(\frac{45}{60} = \frac{3}{4}\right)$, the remainder of each hour being given to relief.

The middle term represents a composite of the last three stages of selection: location within shift which is the second stage, lane within location which is the third stage, and finally, vehicle within lane which is the fourth or ultimate stage of selection. Thus,

$$\frac{1}{24\ fj} = \Pr(\text{Selection within a shift})$$

$$\Pr(\text{Location within shift}) \times \Pr(\text{Lane within Location}) \times \Pr(\text{Vehicle within Lane})$$

For example, at a facility where there are two locations and four lanes are expected to be open during a specified hour at the selected location, the sampling interval for that hour would be determined as follows:

$$\frac{1}{24\ fj} = \left(\frac{1}{2}\right)\left(\frac{1}{4}\right)\left(\frac{1}{K}\right)$$

where K is the sampling interval. Thus,

$$K = \frac{24\ fj}{2 \times 4} = 3\ fj\ .$$

Hence, if fj for this particular shift were 4, the interviewer would select every 12th vehicle.

Locations were selected by random procedures and balanced for relevant factors over each season. Lane rotation patterns at each location within a facility were specified in advance for the interviewer. Fractional intervals K for selection of vehicles within lanes were made into integers by a balanced process of randomization. at all stages of selection random procedures were always carefully specified and used.

It was mentioned previously that the design provided for a self-weighting sample, which is desirable for a number of reasons: (1) Cost of tabulation is reduced; (2) Simplified the analysis of data for meaningful and desirable subclasses; (3) The variance is at a minimum. One reason for adjusting the number of interviews arises from the following facts: (1) The third stage of the sample design is lane within location in which one lane is selected at random from among those open with probability 1/L, where L is the number of lanes assumed open in the location under concern for a specific hour; this number is supplied for each hour operating personal at each facility. (2) The number of lanes actually open at the selected location is noted by the interviewer for the hour interviewing of motorists is taking place.

The adjustment, if it be necessary, comes from the ratio of actual to assumed lanes open and it is needed to provide the correct actual probabilities in the third stage of selection. Then, the over-all probability of selecting a vehicle will remain the same at all facilities, location, etc., and the sample will be self-weighting. (An actual to expected ratio of more than one results in a increase of interviews and

a ratio of less than one in a deficit.)

Similar corrections are made for loss of interviewing time; in very bad weather or in other emergencies the interviewer is permitted to make specified reductions of the working hour and these are noted and adjusted. Non-response also calls for an adjustment, either in case of refusals or if the rushed interviewer must pass a selected vehicle by. Where the need for adjusting the results for the hour arises, this is done by balancing the deficit against possible excess and the adjustment is made by duplicating or omitting cards at random.

Estimates and Sampling Errors

The objectives of the survey consists of many estimates of traffic volume, each estimate for a specified origin-destination couple. Often further specification refers to a defined facility or season, or day or hour, etc. In any case, the estimate is of an aggregate volume of traffic which can be estimated simply as $X'' = Kx$ where K is the inverse of the over-all sampling rate and x is the self-weighting card count in the sample with the specified characteristics. But instead of the above estimator, we use ratio estimators of the form $X' = Nx/n = Np$ where n is the card count for the entire sample or some specified subclass and N is the known total traffic for same; x is defined above. That is, the sample proportion $p = x/n$ for some specified characteristic and subclass is projected to a total by utilizing the information about the aggregate count (N) available from automatic counters.

We used ratio estimates throughout for estimating vehicles or passengers in any characteristic under consideration. They were used because of their ease of application to the data available; because the variance estimates are generally less for ratio estimates than for the simple estimates; and because this eliminates the bias due to the simple "slippage" or "non-coverage" which is bound to creep into field work.

Sampling Error

The two basic types of errors encountered to a greater or lesser degree in all sample surveys affecting the reliability of the results are sampling and non-sampling errors. Non-sampling error stems primarily from errors of response in collecting and in processing the survey results and from any bias in the sample due to non-response. Sampling error arises from the fact that the characteristics as pictured by the sample do not exactly coincide with the characteristics which would emerge from an equal complete coverage of the entire frame.

For computing the sampling error we used a model in which the entire selection for the season consisted of two random and independent halves. (For the sake of simplicity we recommend that this be actually done.) Actually, by "collapsing" strata we created the two computing units; by disregarding some further stratification this results in slight over-estimation of the variance.

The two self-weighting halves for the j-th season, may be represented as follows in estimating the proportion for some characteristic:

$$pj = \frac{x_j}{n_j} = \frac{x_{j1} + x_{j2}}{n_{j1} + n_{j2}}$$

For the entire year of four seasons the similar estimator is the ratio of the sums of the seasons

$$p = \frac{x}{n} = \frac{\sum_{j=1}^{4}(x_{j1} + x_{j2})}{\sum_{j=1}^{4}(n_{j1} + n_{j2})}$$

The "relvariance" (the square of the coefficient of variation) of p can be estimated by:

$$C_p^2 \doteq C_x^2 + C_n^2 - 2\,C\,xn$$

$$\text{where } C_x^2 = \frac{1}{x^2}\sum_{j=1}^{4}(x_{j1} - x_{j2})^2$$

$$C_n^2 = \frac{1}{n^2}\sum_{j=1}^{4}(n_{j1} - n_{j2})^2$$

$$C_{xn} = \frac{1}{xn}\sum_{j=1}^{4}(x_{j1} - x_{j2})\,(n_{j1} - n_{j2})$$

Thus from the sum of four seasonal contrasts for each year's estimates the relvariance can be computed with four degrees of freedom. These computations are made for a large number of items. These are then plotted and average values, subject to smaller variations, are used for estimating standard errors.