# Sampling Methods for Roadside Interviewing

IRWIN MILLER, Graduate Research Assistant, Purdue University;
P.E. IRICK, Assistant Professor of Mathematics, Purdue University;
H.L. MICHAEL, Research Engineer, Joint Highway Research Project, Purdue University;
R.M. BROWN, In Charge Metropolitan Area Traffic Survey,
State Highway Commission of Indiana

PART I. SAMPLING THEORY

PART I of the paper is concerned with an appraisal of the sampling errors in the estimates of the trip frequencies for the various cells of an origin-destination traffic-survey tabulation.

A mathematical exposition is given of the expected errors when the sampling is done at random, by time clusters, and by volume clusters. The paper also discloses the results of an empirical investigation into the sampling errors that actually arose when various sampling methods were applied to the origin-destination tabulations of the Lebanon and Kokomo, Indiana, surveys. These experimental results were in general accord with the theory.

It is concluded that on the average, for the large number of estimates involved in any one survey, the theory of random sampling will satisfactorily explain the errors which arise from the various sampling methods proposed. The results given make it possible to predict the average errors and probability limits for these errors when a particular amount and type of sampling has been done. Conversely, one can determine the amount of sampling necessary to keep the sampling errors within specified probability limits.

PART II. PRACTICAL APPLICATION

IN this second part of the paper, the practical application of sampling to actual field conditions is presented. Station arrangement and sampling procedure for obtaining samples of 50 percent and 25 percent as developed during the conduct of an origin-destination survey at Richmond, Indiana, are discussed. Operational procedure for two sampling methods, volume cluster and time cluster, is given for two sample sizes and for locations having various roadway conditions. Sampling on two-lane, three-lane, and four-lane highways carrying traffic volumes ranging from 1,000 to 12,000 vehicles per day is shown to be both practical and economical. The advantages of each of the two sampling methods for various locations are enumerated.

It is concluded that the use of systematic sampling in the taking of roadside interviews is not only practical and economical but that it has definite operational advantages. The conclusions indicate that a predetermined sampling procedure will place interviewing on a business-like basis that (1) is recognized by the vehicle operator and the community; (2) promotes efficient operation of the interviewing stations; and (3) produces statistically sound results.

● IN recent years, origin-and-destination studies have been of prime importance in the planning, design, and operation of highway facilities. The data from such surveys provide a basis for determining the demand, location, type, and magnitude of new facilities, as well as important facts concerning their operation.

The methods of obtaining the desired data have necessarily been of recent development. Data collection within the survey area has been by means of home interviews, postcards, observations of license plates, route interviews, or combinations of two or more of these. All have been used with varying degrees of success, but only a few actual studies of the relative accuracy and the advantages and disadvantages of each of these methods have been reported. Only the home interview method has received important study toward placing it on a scientific, statistically sound basis (5). One of the purposes of such studies has been directed toward the proper selection of an adequate sample that would be just large enough to produce accurate results in the most economical manner.

Origin-and-destination data of persons entering or leaving the survey area are generally collected by means of roadside interviews at an external cordon. The methods ordinarily used to obtain the necessary information at the cordon result in the selection of a sample which, it is assumed, has been selected in a random manner. In a large majority of the past surveys it is probable that an unbiased sample was selected. A high percentage of the total traffic, however, was often interviewed, so as to minimize the possibilities of obtaining a bias. It is clearly recognized that such almost total sampling is unwarranted if the sample selection is truly random.

The "Manual of Procedures for the Metropolitan Area Traffic Studies" of the Bureau of Public Roads states that "Representative samples of traffic passing each station in both directions should be interviewed." The same manual suggests that in selecting the representative sample:

> No attempt should be made to select for interview a definite predetermined percentage of vehicles passing each station. The number of interviews obtained will depend entirely on the number of interviewers assigned to the station and the number of vehicles passing. The recommended policy is to assign sufficient personnel to each station to interview approximately 50 percent of the traffic passing during peak periods. During the remaining hours of operation the crew will, of course, be able to interview a greater percentage of vehicles passing the station.

The question, however, arises if the selection of such a sample is truly unbiased. Do the interviewers, as good as they may be, have the ability to select an unbiased sample in the field? Do some interviewers select a preponderance of obviously easy interviewees, local vehicles, vehicles containing friends or relatives, or vehicles containing attractive girls? Do some interviewers select too-few foreign vehicles, trucks, or vehicles whose drivers were previously interviewed? Are interviewers prone to pass some vehicles through the station in the interest of less work for themselves? Under such a method, are the vehicles that are interviewed selected randomly over the hour or over that period for which factors are later to be computed?

These questions, undoubtedly, cannot always be truthfully answered in a manner which assures an unbiased sample. Furthermore, the question arises as to what size of sample will give adequate results. Certainly there can be no warrant to take a 75-percent sample if a smaller sample will give the necessary information as accurately as required.

It is generally recognized that a method which would assure the selection of a random sample would be of great value to roadside interviewing. The purpose of this paper is to examine the errors which could be expected from the use of various sampling rates and sampling methods and the practicality of these methods under field conditions.

## PART I. SAMPLING THEORY

Irwin Miller and P. E. Irick

● AT a particular external-survey station, incoming traffic can be considered to originate at the station and have d, say, different destinations. Outgoing traffic destined for this station can have d different origins if d+1 is the total number of internal tracts and external stations. Vehicles passing through a single station can have one of second origin-destination (O. D.) combinations. For any one of these O. D's we define the universe to consist of the N vehicles which could, by virtue of the direction of traffic flow, have had the specified O. D. during the time that the survey station is open. One universe serves for d O. D. cells in the (rectangular) O. D. tabulation. If this universe is sampled in some predetermined manner, the sample estimates of the universe O. D. frequencies are subject to chance variation. It is our purpose to appraise these errors of estimation. We shall first consider the sampling theory that evolves when one makes certain assumptions about the universe and the nature of the sampling. Secondly, we shall show the results of an empirical investigation into these sampling errors.

## RANDOM SAMPLING

Let T be the number of vehicles in a universe of N vehicles which have a particular O.D. If a sample of n vehicles is chosen at random from the N, and t of these have the O.D., then

$$\hat{T} = \frac{N}{n} t \tag{1}$$

is an unbiased estimate of T. The sampling variance of $\hat{T}$ is known (1, p115) to be

$$\text{var}(\hat{T}) = \left(\frac{N - n}{N - 1}\right) \frac{N^2pq}{n} \tag{2}$$

where $p = T/N$, the universe proportion having the O.D., and where $q = 1 - p$. We shall consider the coefficient of variation of $\hat{T}$, $CV(\hat{T})$, to be a measure of the percentage error in $\hat{T}$ which arises because of sampling variation. This measure is defined by

$$CV(\hat{T}) = \frac{[\text{var}(\hat{T})]^{1/2}}{T} \tag{3}$$

where Equation 3 gives the decimal equivalent of the percentage error. Substitution of Equation 2 into 3 gives

$$CV(\hat{T}) = \left[\left(\frac{N - n}{N - 1}\right) \frac{N}{n} \left(\frac{1}{T} - \frac{1}{N}\right)\right]^{1/2} . \tag{4}$$

If the sampling rate is designated by $r=n/N$, and if N is large relative to T, we have the approximate formula

$$CV(\hat{T}) = \left[\left(\frac{1 - r}{r}\right) \frac{1}{T}\right]^{1/2} . \tag{5}$$

Since the universe is not defined until after the period of the survey, there is no opportunity to construct a chance mechanism for the selection of a random sample of n vehicles. One might regard any subset of the universe as being a random sample, but in such a case there would be no assurance that the foregoing formulas were applicable. As soon as one turns to some objective method for selecting the subset, then it becomes necessary to study the sampling variances associated with this method. We shall discuss various cluster sampling methods.

## CLUSTER SAMPLING

### General Theory

Let the universe be divided into G clusters of vehicles, and let the sample consist of g of these clusters. Each cluster is to be completely interviewed, and so there is no sampling variation within clusters. It could be that the g clusters are drawn at random, or that they are chosen systematically, taking every $k^{th}$ universe cluster into the sample. Let $t_i$ be the number of vehicles in the $i^{th}$ cluster ($i = 1, 2, \ldots, G$) which have the specified O.D. and let $t = \sum_{i=1}^{g} t_i$ be the number of vehicles with the O.D. which are found in the sample. If the expected or average value of each $t_i$, with respect to all possible samples of the type selected, is T/G, where $T = \sum_{i=1} t_i$, then

$$\hat{T} = \frac{G}{g} t \tag{6}$$

is an unbiased estimate of T. The sampling variance of $\hat{T}$ is given (1, p. 193) by

$$\text{var}(\hat{T}) = \left(\frac{G - g}{G - 1}\right)\frac{G^2}{g} \text{var}(t_i) \tag{7}$$

where

$$\text{var}(t_i) = \frac{1}{G} \sum_{i=1}^{G}\left(t_i - \frac{T}{G}\right)^2 . \tag{8}$$

The cluster to cluster variance, var $(t_i)$, is a universe parameter peculiar to the universe and the clustering device which has been used. Formula 7 is valid when the g clusters in the sample have been chosen at random from the G available clusters. If the sample clusters are taken systematically (2, p. 157),

$$\text{var}(T) = \left[1 + \frac{2}{g} \sum_{j=1}^{g-1} (g-j)\, \rho_{jk}\right]\frac{G^2}{g} \text{var}(t_i) \tag{9}$$

where $\rho_{jk}$ is the serial correlation coefficient defined by

$$\rho_{jk} = \tag{10}$$

$$\sum_{i=1}^{k(g-j)} \left(t_i - \frac{T}{G}\right)\left(t_{i+jk} - \frac{T}{G}\right)/k(g - j)\, \text{var}(t_i).$$

If the $\rho_{jk}$ values are negative, the implication is that sample clusters with unusually high $t_i$ values will be compensated for by other clusters in the sample with lower than expected $t_i$. In fact, if the $\rho_{jk}$ are sufficiently negative, $var(\hat{T})$ from systematically chosen clusters will be less than $var(\hat{T})$ from randomly chosen clusters. The equality of these two variances occurs when

$$\sum_{j=1}^{g-1} (g-j)\, \rho_{jk} = \frac{-g(g-1)}{2(G-1)} \quad . \qquad (11)$$

There is empirical evidence, to be mentioned later, that Equation 11 may be satisfied on the average for the various O.D. cells in a given survey when particular types of clusters are used. More investigation of this sort should be done, but in the present discussion we shall use Equation 7 to describe the sampling variation in $\hat{T}$. Substituting Equation 7 into 3 we have the expected error for cluster sampling.

$$CV(\hat{T}) = \left[\left(\frac{G-g}{G-1}\right)\frac{G^2}{g}\,\frac{var(t_i)}{T^2}\right]^{1/2} . \qquad (12)$$

TABLE 1

CHARACTERISTICS OF SAMPLING VARIATION IN TIME CLUSTER MODELS

| Time Cluster | G | $E[var(t_i)]$ | $[CV(T)]^2$ g=1 | g=2 | g=3 | 95% limits for $var(t_i)$ | $[CV(T)]^2$ g = 1 | g = 2 | g = 3 |
|---|---|---|---|---|---|---|---|---|---|
| 30 min | 2 | 25T | $\frac{1}{T}$ | | | 96T | $\frac{3\ 89}{T}$ | | |
| 15 min. | 4 | 19T | $\frac{3}{T}$ | $\frac{1}{T}$ | | 49T | $\frac{7\ 81}{T}$ | $\frac{2.60}{T}$ | |
| 10 min | 6 | 14T | $\frac{5}{T}$ | $\frac{2}{T}$ | $\frac{1}{T}$ | .31T | $\frac{11.07}{T}$ | $\frac{4\ 43}{T}$ | $\frac{2.21}{T}$ |

If a sample of g clusters has been taken, $var(t_i)$ can be estimated by

$$var(t_i) = \frac{(G-1)}{G(g-1)} \sum_{i=1}^{g} \left(t_i - \frac{t}{g}\right)^2 . \qquad (13)$$

The question arises as to whether one can make any appraisal of $var(t_i)$ before a sample is taken. This can be done only if one assumes some mathematical model or theoretical mechanism for any particular clustering procedure. One mechanism described in the literature (1, p. 198) assumes that the clusters obtained in any one sample represent a random selection from all possible clusterings of the N universe

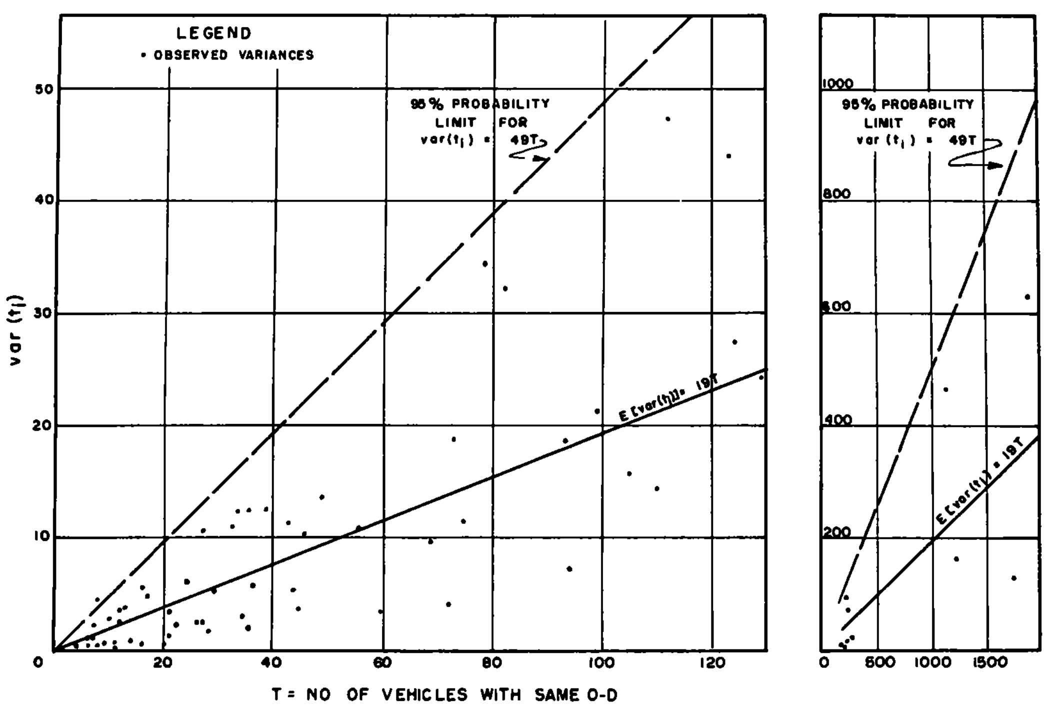


Figure 1. Observed quarter-hour-cluster variances (Lebanon universes).

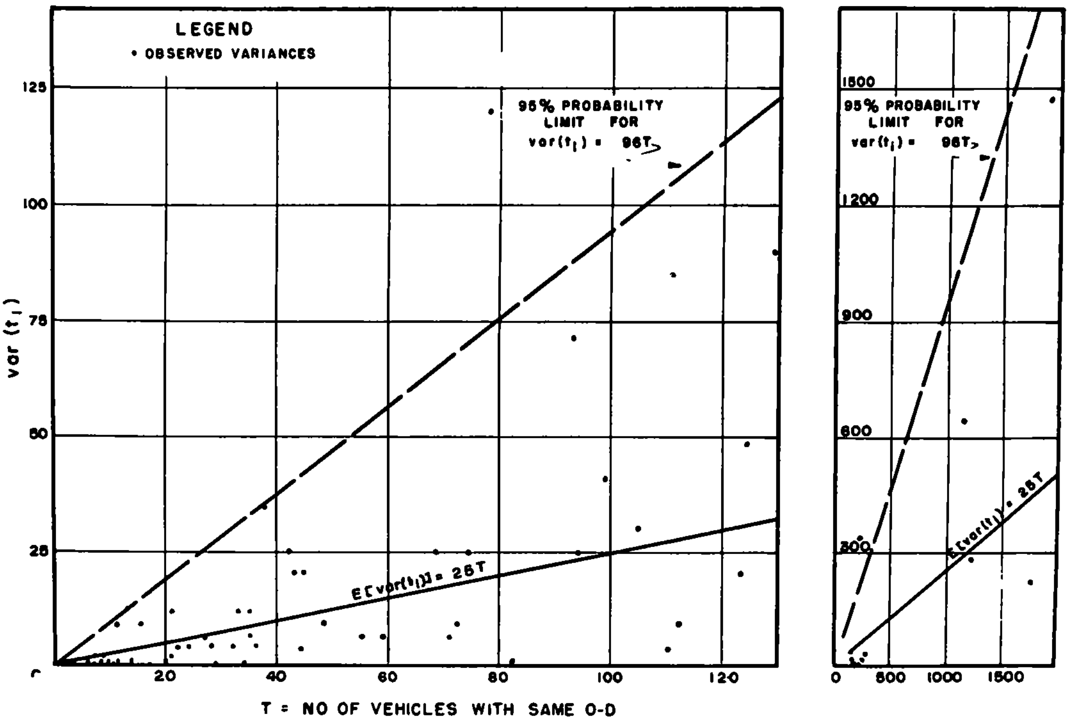


Figure 2. Observed half-hour-cluster variances (Lebanon universes).

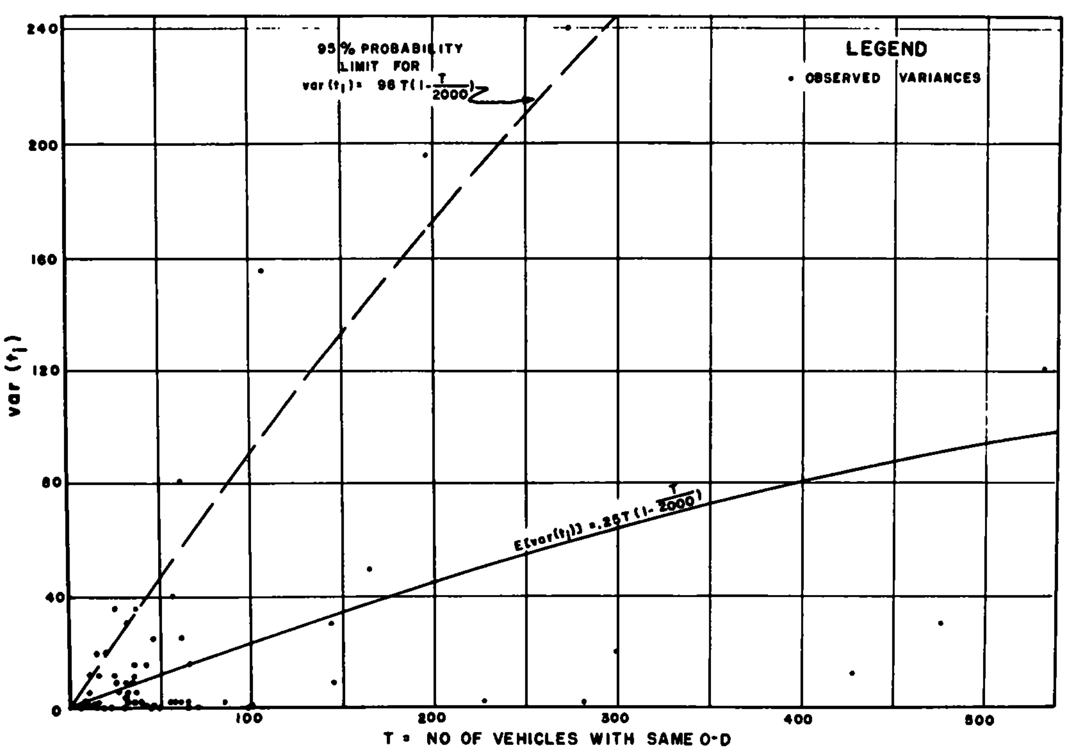


Figure 3. Observed volume-cluster variances (Kokomo universes).

cars into the G clusters. This supposition leads to an average or expected value

$$E\left[\mathrm{var}(t_i)\right]\left(\frac{G-1}{G^2}\right)T\left(1-\frac{T}{N}\right). \qquad (14)$$

We shall discuss two additional models which seem to be intimately related to practical sampling procedures.

Time Clusters

In this method of sampling the survey period is divided into hours, then each hour into the same number of equal time intervals, say 15 min. Let G be the number of intervals in each hour, and $t_i$ be the total number, over all hours, of vehicles which have the O.D. in question and which are found in the $i^{th}$ time interval. Then the extent of sampling variation, where the sample is to consist of g from the G clusters, is dependent upon $\mathrm{var}(t_i)$ as in Equation 7 or 9 depending upon how the g clusters are selected. To estimate $\mathrm{var}(t_i)$ we shall assume that the T cars having the O.D. represent independent "trials" such that there is probability 1/G for each car to fall into any one of the G intervals. Under this assumption it can be shown that the expected value of each $t_i$ is T/G, and from this that Equation 6 gives an unbiased estimate of T. Furthermore, from Equation 8,

$$\mathrm{var}(t_i)=\frac{T}{G^2}\chi^2, \text{ where } \chi^2=\sum_{i=1}^{G}\left(t_i-\frac{T}{G}\right)^2\Big/\frac{T}{G}. \qquad (15)$$

$\chi^2$ follows the chi-squared probability law with G - 1 degrees of freedom. Since the expected value of $\chi^2$ is G - 1, we have

$$E\left[\mathrm{var}(t_i)\right]=\left(\frac{G-1}{G^2}\right)T, \qquad (16)$$

and the corresponding expected error is, from Equations 16 and 12,

$$CV(\hat{T})=\left[\left(\frac{G-g}{g}\right)\frac{1}{T}\right]^{1/2}. \qquad (17)$$

Probability limits for $\mathrm{var}(t_i)$ can be set using tables of chi-square. Table 1 shows several values for Equation 16, upper 95-percent probability limits for $\mathrm{var}(t_i)$, and corresponding values of the squared coefficient of variation for several sample sizes.

Among the various O.D. cells in any one survey, it is almost certain that one will encounter varying degrees of lack of independence of the type assumed. It would seem that any theoretical model can only be true "in the large" and that its appropriateness can best be judged in the light of empirical evidence. Figure 1 shows quarter-hour variances for 68 cells in an O.D. tabulation to be described later (Lebanon). Figure 2 shows the half-hour variances for the same cells. The solid and dotted lines are plotted from expected values and probability limits given in Table 1. The serial correlations for the quarter hour clusters in these cells varied considerably, but averaged to be -0.33, which happens to be the value specified by Equation 11.

Tables 2 and 3 give the data for Figures 1 and 2 respectively. The $t_i$ values are vehicle counts over a 12-hr. period of time.

Volume Clusters

Suppose that the universe of N cars is divided into consecutive sets of c cars each so that there are N/c (approximately) such sets. We may take c to be the number of interviewers on duty at the survey station. Let g such sets be interviewed, then G - g sets waved through the station without being interviewed, etc. Then the universe has been clustered into G clusters, g of which appear in the sample. The $i^{th}$ such cluster has $t_i$ vehicles which have the O-D in question and $(N/G)-t_i$ which do not. We assume that, for every vehicle, the probability of having the O-D, p=T/N, is independent of the probability of being in any particular cluster, and that the latter probability is 1/G. This model then describes a 2 x G contingency table (2, p. 227) and, under the assumptions,

$$\chi^2=\frac{G^2}{qT}\mathrm{var}(t_i) \qquad (18)$$

has nearly the probability distribution of chi-squared with G - 1 degrees of freedom. The expected value of $\mathrm{var}(t_i)$ is as in Equation 14, and using this average value in Equation 12 we obtain

$$CV(\hat{T})=\left[\left(\frac{G-g}{g}\right)\left(\frac{1}{T}-\frac{1}{N}\right)\right]^{1/2} \qquad (19)$$

where $\hat{T}$ is calculated from Equation 6. Again we have considered the systematic

TABLE 2

QUARTER HOUR VARIANCES - LEBANON DATA

| T | $t_1$ | $t_2$ | $t_3$ | $t_4$ | var($t_1$) | ρ 12 | T | $t_1$ | $t_2$ | $t_3$ | $t_4$ | var($t_1$) | ρ 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 1 | 0 | 1 | 2 | 5 | -1 00 | 36 | 11 | 5 | 10 | 10 | 5 5 | - 18 |
| 6 | 1 | 1 | 3 | 1 | 5 | 00 | 38 | 6 | 7 | 10 | 15 | 12 2 | - 63 |
| 6 | 1 | 2 | 0 | 3 | 1 2 | 60 | 42 | 5 | 11 | 14 | 12 | 11 2 | - 82 |
| 7 | 2 | 0 | 3 | 2 | 1 2 | - 05 | 43 | 14 | 12 | 10 | 7 | 5 1 | - 58 |
| 7 | 1 | 2 | 4 | 0 | 2 2 | - 49 | 44 | 13 | 11 | 8 | 12 | 3 5 | - 85 |
| 8 | 2 | 2 | 3 | 1 | 5 | 00 | 45 | 11 | 16 | 7 | 11 | 10 2 | - 06 |
| 8 | 0 | 3 | 0 | 5 | 4 5 | 78 | 48 | 15 | 12 | 15 | 6 | 13 5 | 33 |
| 9 | 3 | 3 | 2 | 1 | 7 | - 82 | 55 | 19 | 11 | 14 | 11 | 10 7 | 41 |
| 10 | 1 | 3 | 5 | 1 | 2 8 | - 82 | 59 | 15 | 12 | 15 | 17 | 3 2 | - 87 |
| 11 | 3 | 3 | 3 | 2 | 2 | - 33 | 68 | 19 | 20 | 12 | 17 | 9 5 | - 53 |
| 11 | 2 | 3 | 2 | 4 | 7 | 63 | 71 | 17 | 21 | 17 | 16 | 3 7 | - 69 |
| 12 | 2 | 1 | 6 | 3 | 3 5 | - 42 | 72 | 22 | 11 | 18 | 21 | 18 5 | - 46 |
| 12 | 5 | 2 | 4 | 1 | 2 5 | 80 | 74 | 15 | 17 | 24 | 18 | 11 2 | - 82 |
| 13 | 4 | 6 | 1 | 2 | 3 7 | - 69 | 78 | 16 | 12 | 23 | 27 | 34 2 | -1 00 |
| 14 | 4 | 4 | 2 | 4 | 8 | - 33 | 82 | 20 | 20 | 13 | 29 | 32 2 | - 01 |
| 16 | 8 | 3 | 2 | 3 | 5 5 | - 64 | 93 | 20 | 18 | 27 | 28 | 18 7 | - 99 |
| 16 | 4 | 4 | 3 | 5 | 5 | 00 | 94 | 27 | 25 | 22 | 20 | 7 2 | - 72 |
| 17 | 4 | 5 | 7 | 1 | 4 7 | - 33 | 99 | 17 | 26 | 29 | 27 | 21 2 | - 71 |
| 20 | 5 | 4 | 5 | 6 | 5 | -1 00 | 105 | 31 | 27 | 20 | 27 | 15 7 | - 93 |
| 21 | 5 | 4 | 5 | 7 | 1 2 | - 89 | 110 | 31 | 22 | 26 | 31 | 14 2 | - 86 |
| 21 | 4 | 3 | 7 | 7 | 3 2 | - 96 | 112 | 30 | 29 | 36 | 17 | 47 5 | 05 |
| 22 | 7 | 6 | 3 | 6 | 2 2 | - 78 | 123 | 35 | 22 | 39 | 27 | 44 2 | 77 |
| 22 | 8 | 5 | 4 | 5 | 2 2 | - 76 | 124 | 28 | 27 | 40 | 29 | 27 5 | - 35 |
| 24 | 10 | 4 | 4 | 6 | 6 0 | - 67 | 129 | 26 | 29 | 38 | 36 | 24 2 | - 99 |
| 26 | 8 | 5 | 8 | 5 | 2 2 | 1 00 | 195 | 47 | 47 | 47 | 54 | 9 2 | - 67 |
| 27 | 9 | 7 | 5 | 6 | 2 2 | - 94 | 198 | 46 | 52 | 48 | 52 | 6 8 | 85 |
| 27 | 7 | 4 | 12 | 4 | 10 7 | 41 | 203 | 54 | 47 | 56 | 46 | 18 7 | - 02 |
| 28 | 8 | 8 | 5 | 7 | 1 5 | - 67 | 205 | 64 | 57 | 38 | 46 | 99 7 | -1 00 |
| 29 | 5 | 10 | 5 | 9 | 5 2 | 95 | 226 | 62 | 65 | 66 | 43 | 76 2 | 21 |
| 32 | 11 | 3 | 7 | 11 | 11 0 | - 82 | 236 | 53 | 59 | 58 | 66 | 21 5 | 30 |
| 33 | 14 | 6 | 8 | 5 | 12 2 | 24 | 1111 | 303 | 278 | 286 | 244 | 461 2 | 22 |
| 34 | 7 | 9 | 7 | 11 | 2 8 | 64 | 1218 | 303 | 323 | 305 | 287 | 162 8 | -1 00 |
| 35 | 12 | 9 | 11 | 3 | 12 2 | 24 | 1758 | 425 | 429 | 457 | 437 | 130 8 | - 97 |
| 35 | 10 | 10 | 7 | 8 | 1 7 | - 92 | 1895 | 478 | 508 | 472 | 437 | 636 2 | -1 00 |

TABLE 3

HALF HOUR VARIANCES - LEBANON DATA

| T | $t_1$ | $t_2$ | var($t_1$) | T | $t_1$ | $t_2$ | var(t1) |
|---|---|---|---|---|---|---|---|
| 4 | 1 | 3 | 1.0 | 36 | 16 | 20 | 4 0 |
| 6 | 2 | 4 | 1 0 | 38 | 13 | 25 | 36 0 |
| 6 | 3 | 3 | 0 | 42 | 16 | 26 | 25.0 |
| 7 | 2 | 5 | 2 2 | 43 | 26 | 17 | 20 2 |
| 7 | 3 | 4 | .2 | 44 | 24 | 20 | 4 0 |
| 8 | 4 | 4 | 0 | 45 | 27 | 18 | 20 2 |
| 8 | 3 | 5 | 1.0 | 48 | 27 | 21 | 9 0 |
| 9 | 6 | 3 | 2 2 | 55 | 30 | 25 | 6 2 |
| 10 | 4 | 6 | 1.0 | 59 | 27 | 32 | 6 2 |
| 11 | 6 | 5 | .2 | 68 | 39 | 29 | 25 0 |
| 11 | 5 | 6 | 2 | 71 | 38 | 33 | 6 2 |
| 12 | 3 | 9 | 9 0 | 72 | 33 | 39 | 9 0 |
| 12 | 7 | 5 | 1.0 | 74 | 32 | 42 | 25 0 |
| 13 | 10 | 3 | 12 2 | 78 | 28 | 50 | 121.0 |
| 14 | 8 | 6 | 1 0 | 82 | 40 | 42 | 1 0 |
| 16 | 11 | 5 | 9.0 | 93 | 38 | 55 | 72 2 |
| 16 | 8 | 8 | .0 | 94 | 52 | 42 | 25.0 |
| 17 | 9 | 8 | .2 | 99 | 43 | 56 | 42 2 |
| 20 | 9 | 11 | 1 0 | 105 | 58 | 47 | 30.2 |
| 21 | 9 | 12 | 2 2 | 110 | 53 | 57 | 4.0 |
| 21 | 7 | 14 | 12 2 | 112 | 59 | 53 | 9 0 |
| 22 | 13 | 9 | 4.0 | 123 | 57 | 66 | 20 2 |
| 22 | 13 | 9 | 4 0 | 124 | 55 | 69 | 49.0 |
| 24 | 14 | 10 | 4 0 | 129 | 55 | 74 | 90.2 |
| 26 | 13 | 13 | 0 | 195 | 94 | 101 | 12 2 |
| 27 | 16 | 11 | 6 2 | 198 | 98 | 100 | 1 0 |
| 27 | 11 | 16 | 6.2 | 203 | 101 | 102 | .2 |
| 28 | 16 | 12 | 4 0 | 205 | 121 | 84 | 342 2 |
| 29 | 15 | 14 | .2 | 226 | 117 | 109 | 16 0 |
| 32 | 14 | 18 | 4.0 | 236 | 112 | 124 | 36 0 |
| 33 | 20 | 13 | 12.2 | 1111 | 581 | 530 | 650 2 |
| 34 | 16 | 18 | .2 | 1218 | 626 | 592 | 289.0 |
| 35 | 21 | 14 | 12 2 | 1758 | 864 | 894 | 225.0 |
| 35 | 20' | 15 | 6.2 | 1895 | 986 | 909 | 1482 2 |

selection of the g clusters to be equivalent to the corresponding random selection. Comparison of Equation 19 with 17 and Equation 18 with 15 shows that Table 1 can be used for this volume cluster model provided the factor $q = 1 - \frac{T}{N}$ is inserted as a multiplier for each entry there. Figure 3 and the accompanying data in Table 4 shows 68 observed values of var($t_1$) from Kokomo universes where G = 2, g = 1, N = 2,000 (approx.). These variances, however, are not from true volume clusters as will be pointed out later. Table 1, modified as just stated, gives the solid and dotted curves in Figure 3.

In summary it should be pointed out that Formula 14 substituted into Equations 12, 16, and 19 are very similar, and in fact, that all three are practically equivalent to Formulas 4 and 5. This is so because the assumptions made for cluster sampling have led to the conclusion that we shall expect cluster sampling errors to average to be the same as random sampling errors even though the two types of errors may be different for a single universe. Having discussed the theoretical sampling variation to be expected through the use of these methods, we now turn our attention to the observed sampling errors which arose when these methods were applied.

## DESCRIPTION OF EMPIRICAL INVESTIGATIONS

The results of census surveys taken at Lebanon and Kokomo, Indiana, comprise the universes from which samples were drawn in an empirical study of sampling variances. It was impracticable to interview every vehicle when traffic was unusually heavy, but over 90 percent of the traffic was interviewed in each survey, making it reasonable to regard these surveys as complete. Table 5 shows the O. D. frequencies for the Lebanon survey.

Traffic at Lebanon was interviewed at eight different external stations, and the area was divided into 24 internal tracts. Vehicles passing through Station 8, for example, could have had 24 internal-external O. D. combinations, and 14 through-trip combinations (7 inbound and 7 outbound) or 62 in all. For the O.D. Cell 8-010, the universe consists of N = 4703.4 vehicles inbound through Station 8. For the O-D Cell 8-3 the universe has been taken to consist of the 4,703.4 vehicles which went through Station 8 in the inbound direction, plus the 947.7 vehicles passing through Station 3 in the outbound direction, or 5,651.1 vehicles in all. As can be seen in Table 5, the respective T values for these cells are 91.6, and 184.8. These values are not actual counts, since the counts were adjusted for day to day and other variations by means of several factors (3).

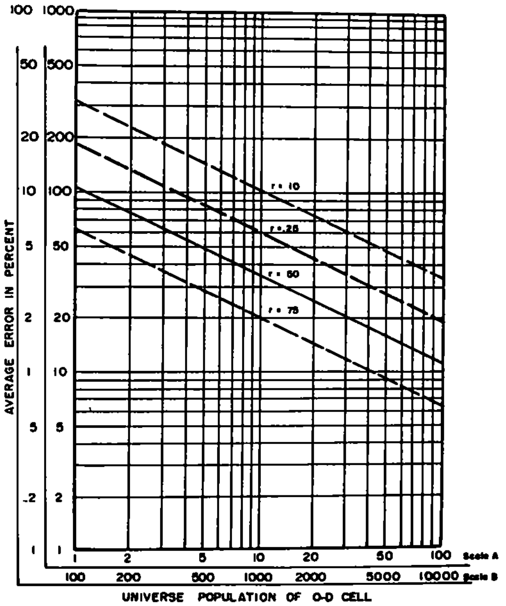


Figure 4. Expected errors due to uniform sampling (for various sampling rates).

## Types of Samples

The data for each universe interview was recorded on a punch card. The card showed the quarter hour of the hour in which the interview was taken and serial numbers were assigned to the cards within the quarter hours. However, these serial numbers were not necessarily in the order of occurrence of the interview, due to the fact that several interviews were being conducted simultaneously. Consequently, time-cluster samples were limited to rates of 25, 50, and 75 percent, and it was felt reasonable only to consider the volume cluster sample which compares even and odd serial numbers. Such rates are de-

TABLE 4

VOLUME CLUSTER VARIANCES - KOKOMO DATA

| T | $t_1$ | $t_2$ | var($t_i$) | T | $t_1$ | $t_2$ | var($t_i$) |
|---|---|---|---|---|---|---|---|
| 6 | 2 | 4 | 1.0 | 39 | 21 | 18 | 2.2 |
| 7 | 3 | 4 | .2 | 40 | 19 | 21 | 1.0 |
| 7 | 3 | 4 | 2 | 42 | 20 | 22 | 1.0 |
| 8 | 5 | 3 | 1.0 | 42 | 20 | 22 | 1.0 |
| 9 | 6 | 3 | 2.2 | 42 | 17 | 25 | 16.0 |
| 10 | 5 | 5 | 0 | 45 | 23 | 22 | .2 |
| 10 | 4 | 6 | 1.0 | 46 | 28 | 18 | 25.0 |
| 11 | 8 | 3 | 6.2 | 46 | 22 | 24 | 1.0 |
| 11 | 2 | 9 | 12.2 | 49 | 24 | 25 | .2 |
| 14 | 8 | 6 | 1.0 | 55 | 29 | 26 | 2.2 |
| 15 | 12 | 3 | 20.2 | 57 | 19 | 38 | 40.2 |
| 15 | 8 | 7 | .2 | 57 | 30 | 27 | 2.2 |
| 17 | 10 | 7 | 2.2 | 61 | 29 | 32 | 2.2 |
| 17 | 12 | 5 | 12.2 | 62 | 40 | 22 | 81.0 |
| 18 | 9 | 9 | .0 | 62 | 36 | 26 | 25.0 |
| 19 | 10 | 9 | 2 | 65 | 33 | 32 | .2 |
| 21 | 6 | 15 | 20.2 | 65 | 31 | 34 | 2.2 |
| 23 | 12 | 11 | .2 | 66 | 29 | 37 | 16.0 |
| 25 | 16 | 9 | 12.2 | 71 | 35 | 36 | .2 |
| 26 | 10 | 16 | 9.0 | 85 | 44 | 41 | 2.2 |
| 26 | 7 | 19 | 36.0 | 97 | 48 | 49 | 2 |
| 27 | 11 | 16 | 6.2 | 100 | 49 | 51 | 1.0 |
| 29 | 14 | 15 | 2 | 107 | 41 | 66 | 156.2 |
| 30 | 18 | 12 | 9.0 | 143 | 77 | 66 | 30.2 |
| 30 | 13 | 17 | 4.0 | 144 | 69 | 75 | 9.0 |
| 31 | 14 | 17 | 2.2 | 164 | 89 | 75 | 49.0 |
| 31 | 10 | 21 | 30.2 | 196 | 112 | 84 | 196.0 |
| 31 | 13 | 18 | 6.2 | 227 | 112 | 115 | 2.2 |
| 34 | 20 | 14 | 9.0 | 273 | 152 | 121 | 240.2 |
| 35 | 14 | 21 | 12.2 | 281 | 139 | 142 | 2.2 |
| 35 | 16 | 19 | 2.2 | 299 | 145 | 154 | 20.2 |
| 36 | 14 | 22 | 16.0 | 427 | 210 | 217 | 12.2 |
| 37 | 24 | 13 | 36.0 | 475 | 243 | 232 | 30.2 |
| 37 | 21 | 16 | 6.2 | 532 | 255 | 277 | 121.0 |

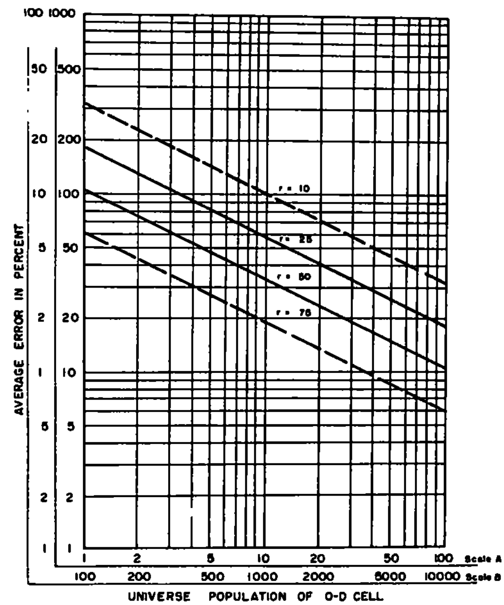


Figure 5. Expected errors due to quarter-hour sampling.

TABLE 5

ORIGIN AND DESTINATION UNIVERSE TABULATION OF NUMBER OF VEHICULAR TRIPS FOR LEBANON BY-PASS "BEFORE" STUDY

External Stations - Origins

| Internal Tracts - Destinations | 1 | 3 | 5 | 8 | 10 | 11 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|
| 001 | 3 7 | 1 1 | 1 1 | 2 2 | 1 5 | 1 3 | 0 0 | 0 0 |
| 002 | 16 7 | 2 3 | 1 3 | 15 3 | 1 2 | 0 0 | 2.2 | 6 3 |
| 003 | 6 7 | 4 7 | 11.1 | 24 9 | 5 3 | 4.0 | 1.3 | 10.9 |
| 004 | 16 5 | 17 7 | 29 2 | 40 4 | 23 3 | 10 1 | 4 9 | 35 0 |
| 005 | 5 5 | 2 4 | 10 2 | 24 8 | 5 9 | 10 0 | 9.2 | 15 1 |
| 006 | 2 4 | 12 9 | 11 3 | 30 9 | 14 5 | 7 8 | 11 7 | 18 9 |
| 007 | 1 0 | 21.1 | 15 6 | 9 0 | 5.6 | 0.0 | 0 0 | 1 0 |
| 008 | 39 4 | 40 9 | 62 3 | 51 6 | 13 2 | 22 5 | 14.3 | 59 8 |
| 009 | 8 1 | 17 1 | 19 4 | 42 4 | 6 2 | 7 4 | 12 8 | 24 7 |
| 010 | 41 1 | 50 9 | 57 0 | 91 6 | 33 3 | 42.7 | 26 6 | 67 1 |
| 011 | 2 4 | 3 9 | 0 9 | 14 1 | 7.1 | 3 0 | 1 8 | 12.2 |
| 012 | 8.5 | 21.4 | 17 0 | 32 3 | 6 1 | 3 2 | 3 3 | 18 8 |
| 013 | 4 1 | 19.0 | 29 7 | 11 9 | 5 6 | 7 3 | 7 2 | 18 1 |
| 014 | 20 8 | 49.4 | 48 5 | 71 1 | 11 7 | 13 9 | 7 9 | 30 5 |
| 015 | 163 2 | 193.9 | 203 0 | 278 7 | 97.3 | 100 3 | 74 5 | 167 6 |
| 016 | 12 9 | 26 5 | 76.3 | 39 9 | 18 2 | 29 7 | 29 8 | 33.3 |
| 017 | 16 8 | 16 0 | 28 4 | 84 9 | 18.5 | 36 9 | 15 6 | 27 3 |
| 018 | 2 2 | 2.8 | 11 0 | 7 2 | 1.1 | 0 0 | 0 0 | 0.0 |
| 019 | 15 1 | 17 9 | 27 0 | 56 5 | 18 4 | 19 0 | 6 0 | 23 8 |
| 020 | 6 5 | 3 7 | 4.5 | 29 2 | 5 3 | 1 1 | 4 7 | 16 5 |
| 021 | 9 1 | 20 8 | 33 0 | 46 3 | 18 5 | 16 7 | 10 9 | 35 8 |
| 022 | 1 3 | 1 0 | 16 5 | 6 2 | 1 3 | 0 0 | 0 0 | 2 9 |
| 023 | 2 3 | 2 9 | 6.3 | 19 9 | 6 7 | 5 4 | 18 4 | 6 5 |
| 024 | 1 1 | 1 3 | 2 6 | 20 5 | 9 7 | 2 7 | 0 0 | 3 8 |

External Stations - Destinations

| Internal Tracts - Origins | 1 | 3 | 5 | 8 | 10 | 11 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|
| 001 | 5.4 | 0 0 | 1 2 | 4 4 | 0 0 | 0.0 | 0.0 | 0 0 |
| 002 | 7 3 | 4 5 | 1 5 | 16 6 | 4 5 | 2 4 | 1.1 | 12.0 |
| 003 | 10.2 | 9.6 | 11 7 | 30 4 | 3.0 | 4.5 | 3 7 | 9 8 |
| 004 | 15 9 | 19 9 | 32 3 | 34 5 | 16 9 | 6 5 | 4.3 | 18.4 |
| 005 | 6 1 | 1.1 | 5 5 | 32 5 | 4.4 | 2.0 | 11.9 | 9.2 |
| 006 | 5 5 | 10.4 | 9.8 | 32.2 | 8.0 | 9.3 | 9.3 | 29 0 |
| 007 | 4 3 | 7.8 | 10 0 | 2.4 | 3.1 | 0.0 | 0 0 | 1.1 |
| 008 | 49 9 | 55 9 | 59 2 | 54 8 | 15 9 | 15.6 | 17.8 | 54 3 |
| 009 | 14.5 | 21.5 | 20 0 | 48 9 | 10 5 | 3 4 | 11 7 | 24.8 |
| 010 | 36 1 | 48.9 | 76 1 | 83 7 | 30 6 | 21 0 | 20 9 | 68 9 |
| 011 | 5.4 | 4.3 | 0.0 | 13 9 | 5.5 | 5.1 | 2.7 | 13 0 |
| 012 | 11 1 | 27 7 | 13 9 | 25 8 | 7 1 | 1.7 | 3 6 | 21 8 |
| 013 | 4 9 | 23 5 | 25 4 | 10 4 | 7 3 | 7 4 | 5 0 | 22.2 |
| 014 | 25.2 | 37.8 | 40.9 | 63 9 | 14 9 | 11 5 | 12 2 | 31 7 |
| 015 | 134 1 | 129 9 | 199 5 | 242 5 | 78.1 | 104 8 | 72 1 | 47 9 |
| 016 | 14 3 | 26 8 | 61.4 | 41 6 | 31 8 | 38 1 | 25.0 | 29 1 |
| 017 | 17 4 | 26 9 | 29 0 | 77 4 | 25 4 | 23 3 | 19 7 | 36 0 |
| 018 | 4 1 | 1 0 | 5 3 | 7 7 | 3 8 | 0.0 | 0.0 | 2.3 |
| 019 | 15 5 | 22 4 | 34 2 | 52 3 | 19 3 | 17 8 | 11 4 | 32 2 |
| 020 | 7 4 | 1 3 | 6 7 | 23 7 | 9.3 | 1.2 | 2 2 | 11 8 |
| 021 | 16 2 | 14 6 | 30 3 | 41 9 | 18 8 | 17.9 | 12 7 | 40 3 |
| 022 | 3 0 | 0 0 | 8 7 | 9.2 | 1.3 | 0 0 | 0 0 | 1 2 |
| 023 | 1 0 | 2.3 | 6 9 | 19 0 | 5 5 | 5 2 | 13 9 | 4 6 |
| 024 | 4 0 | 0 0 | 3 0 | 25.5 | 2.5 | 0 0 | 0 0 | 3 2 |

External Stations - Destinations

| External Stations - Origins | 1 | 3 | 5 | 8 | 10 | 11 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|
| 1 | | 12 8 | 67 8 | 4175 6 | 102 8 | 10.6 | 3 6 | 0 0 |
| 3 | 4.4 | | 26 4 | 206 8 | 220 2 | 7 6 | 6 6 | 31 6 |
| 5 | 62.0 | 27 2 | | 38 4 | 65 6 | 14 4 | 15 6 | 111 0 |
| 8 | 3949 6 | 184 8 | 25 2 | | 29 2 | 8 4 | 8 0 | 498 2 |
| 10 | 132 4 | 168.8 | 62 6 | 30 0 | | 3 2 | 1 2 | 198 4 |
| 11 | 13 2 | 4 6 | 9 4 | 11 4 | 2 6 | | 0.0 | 12 0 |
| 13 | 0.8 | 7.2 | 18 0 | 9 0 | 1 2 | 1 2 | | 1 0 |
| 14 | 0 0 | 44 2 | 131 6 | 523 8 | 218 0 | 8 0 | 2.4 | |

sirable, however, in that they permit reasonable frequencies to appear in cells with small T values. The following samples were selected for study.

Sample 1 is a volume cluster sample, each cluster consisting of $\frac{N}{2}$ vehicles. The sample selected represented the even serial numbers, and thus is a 50-percent sample with G = 2, and g = 1.

Sample 2 is a quarter-hour time-cluster sample. A 50-percent sample was obtained by choosing vehicles which were interviewed in alternate directions by quarter-hour time intervals. Thus, G = 4 and g = 2 in this case.

Sample 3 differs from Sample 2 only in the use of a half-hour time interval, or in that G = 2 and g = 1. Both Samples 2 and 3 would be expected to provide 50-percent samples, since the inbound universe was sampled in the periods during which the outbound universe was not sampled, and vice versa.

Sample 4 combines both the volume cluster technique of sample 1 and the time cluster sampling of sample 3. In this sample the even numbered vehicles were interviewed every other half hour, yielding an expected 25 percent sample.

## Sampling Errors

After a drawing of any of these samples, $\hat{T}$ was computed from Equation 6 for each cell, and compared with the corresponding universe frequency, T. The relative error of estimate in this cell is given by the formula

$$e = \frac{|\hat{T} - T|}{T}. \qquad (20)$$

Table 6 shows the relative errors in percent arising from Sample 2 drawn from the Lebanon universes. Similar tables have been constructed for the remaining samples from both surveys (4).

The variance of $t_i$ has been computed directly from the universe data in the case of cluster sampling. Graphs of these values versus T are typified by Figures 1, 2, and 3, and show that $var(t_i)$ is approximately proportional to T. Then it follows from the theory that

$$CV(\hat{T}) = \sqrt{\frac{K}{T}} \qquad (21)$$

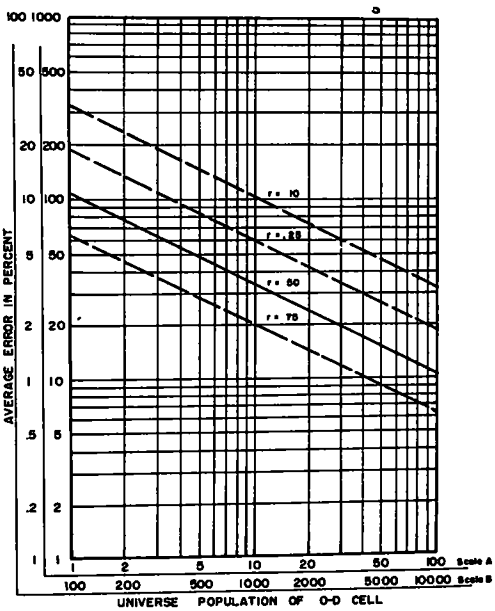


Figure 6. Expected errors due to half-hour sampling.

where K is a constant which will be left for the data itself to determine. In order to estimate K, let us first make the following change of variables. Let

$$Y = \left[CV(\hat{T})\right]^2 ; \quad X = \frac{1}{T} . \tag{22}$$

When this is done, Equation 21 becomes

$$Y = KX. \tag{23}$$

Now each complete sample, as illustrated in Table 6, gives rise to s numbers $e_j$, one for each O.D. cell, and the universe tabulation (e.g., Table 5) provides s corresponding values $T_j$. (The subscripts j are used to indicate that these values came from the $j^{th}$ O.D. cell). Since $e_j$ is the only available estimate of $CV(\hat{T})$ in the $j^{th}$ cell, we shall compute $Y_j' = e_j^2$ our estimate of $Y_j$, from this number. K is estimated by the method of least squares, using the formula

$$\hat{K} = \sum_{j=1}^{s} X_j Y_j' \Big/ \sum_{j=1}^{s} X_j^2 . \tag{24}$$

Under the assumption that the errors arising in the $j^{th}$ cell from all possible samples of the type selected are normally distributed with mean zero and standard deviation $CV(\hat{T}_j)$, it is found (4) that the variance of $\hat{K}$ can be estimated with the aid of the statistic

$$\text{Var } \hat{K} = 2 \hat{K}^2 \sum_{j=1}^{s} X_j^4 \Big/ \left( \sum_{j=1}^{s} X_j^2 \right)^2 . \tag{25}$$

This normality assumption has been validated (4) from the observed data for cells having more than 25 vehicles in the universe. We should not expect normal distribution of errors for the smaller cells.

Computation of $\hat{K}$ together with its variance affords a way to evaluate the results of the plans, as well as a means for estimating $var(t_1)$. Table 7 exhibits these results for the four samples under investigation with respect to the Lebanon and Kokomo universes. The computations have been carried out for Sample 4 in the manner outlined, although no theory is presented here for this type of sample. The general theory related to this method appears in the literature (1, p. 144).

From Table 1 it can be noted that $\hat{K}$ values for samples 2 and 3 were expected to have been 1.00 with upper probability limits of 2.60 and 3.89 respectively. Table 7 indicates that the corresponding average empirical values for $\hat{K}$ were 1.18 and 1.34. The magnitude of the standard errors of these $\hat{K}$ values implies that the empirical estimates of K might vary considerably with more experimentation of the type done here, but also makes it credible that the observed sampling errors are fairly well in accordance with the theory which resulted in Table 1. On the other hand, since all 4 $\hat{K}$ values from Samples 2 and 3 were greater than 1,

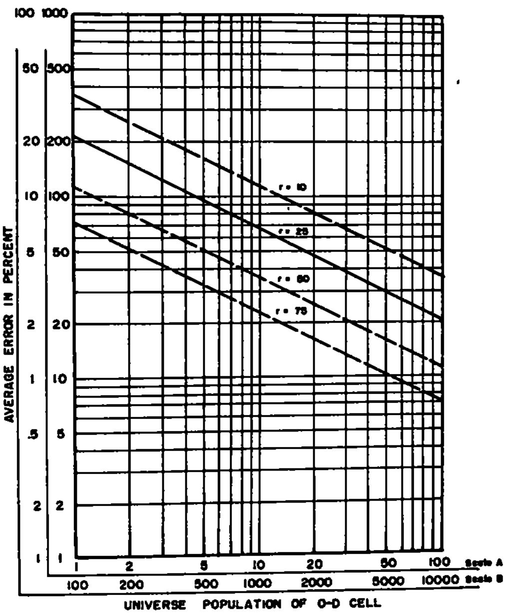


Figure 7. Expected errors due to two-stage sampling.

TABLE 6

OBSERVED ERRORS (IN PERCENT) OF ORIGIN AND DESTINATION TABULATION
SAMPLE 2 FOR LEBANON BY-PASS "BEFORE" STUDY

| Internal Tracts - Destinations | External Stations - Origins | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 3 | 5 | 8 | 10 | 11 | 13 | 14 |
| 001 | 100 0 | 100 0 | 120 0 | 9 1 | 100 0 | 100 0 | 0 0 | 0 0 |
| 002 | 9 8 | 100 0 | 116 7 | 45 3 | 100 0 | 0 0 | 100 0 | 71 0 |
| 0Q3 | 6 1 | 47 8 | 9 1 | 1 6 | 5 0 | 35 0 | 116 7 | 64 8 |
| 004 | 28 4 | 16 1 | 11 1 | 20 1 | 17 4 | 48 0 | 12 5 | 4 0 |
| 005 | 63 0 | 8 3 | 72 0 | 13 1 | 17 2 | 10 2 | 44 4 | 34 7 |
| 006 | 8 3 | 26 6 | 51 8 | 6 6 | 16 7 | 21 1 | 3 4 | 29 0 |
| 007 | 100 0 | 36 5 | 61 0 | 77 3 | 28 6 | 0 0 | 0.0 | 100 0 |
| 008 | 47 9 | 24 8 | 12 1 | 12 2 | 10 8 | 46 8 | 22 5 | 6 1 |
| 009 | 12 5 | 47 6 | 10 4 | 2 4 | 58 1 | 70.3 | 27 0 | 16 4 |
| 010 | 12 3 | 8 0 | 11 0 | 2 9 | 29 3 | 8 1 | 9 2 | 10 9 |
| 011 | 8 3 | 26.3 | 100 0 | 8.6 | 2 9 | 100 0 | 11 1 | 5 0 |
| 012 | 21 4 | 26 4 | 47 6 | 42 1 | 16 7 | 100 0 | 18 8 | 24 7 |
| 013 | 10 0 | 22.3 | 22 4 | 27 1 | 17 9 | 47 2 | 38.9 | 20 2 |
| 014 | 30 1 | 20 9 | 13 8 | 4 8 | 39 7 | 15 9 | 12 8 | 32 7 |
| 015 | 19.1 | 6 0 | 1 2 | 5 4 | 7 5 | 17 0 | 16 8 | 5 4 |
| 016 | 21 9 | 29 0 | 18 4 | 35 5 | 7 8 | 2 7 | 6 8 | 0 6 |
| 017 | 7 2 | 31.6 | 12 9 | 12 4 | 19 8 | 29 7 | 10 4 | 10 4 |
| 018 | 9 1 | 100 0 | 48.1 | 61.1 | 100 0 | 0 0 | 0 0 | 0 0 |
| 019 | 20 0 | 2 3 | 9 0 | 19 7 | 22 0 | 5.3 | 63 3 | 11.1 |
| 020 | 71 9 | 100 0 | 40 9 | 16 0 | 69 2 | 100 0 | 8 7 | 19 8 |
| 021 | 2 2 | 30.1 | 12 3 | 13 2 | 15 4 | 73 2 | 20 4 | 25 4 |
| 022 | 100 0 | 100 0 | 42 0 | 16 1 | 116 7 | 0 0 | 0 0 | 35.7 |
| 023 | 18 2 | 107 1 | 22 6 | 10 2 | 39 4 | 51 9 | 5 5 | 28 1 |
| 024 | 100.0 | 116 7 | 100 0 | 34 7 | 27 1 | 23 1 | 0 0 | 36 8 |

| Internal Tracts - Origins | External Stations - Destinations | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 3 | 5 | 8 | 10 | 11 | 13 | 14 |
| 001 | 107 7 | 0 0 | 0.0 | 109 5 | 0 0 | 0.0 | 0.0 | 0.0 |
| 002 | 5 1 | 104 5 | 100 0 | 26 1 | 54.5 | 0.0 | 0 0 | 10 3 |
| 003 | 10 9 | 6 4 | 6 5 | 5 8 | 100 0 | 9 1 | 100.0 | 10 1 |
| 004 | 11 5 | 0 9 | 11 2 | 11 9 | 9 2 | 25 0 | 42 9 | 16 5 |
| 005 | 36 1 | 120 0 | 44 4 | 2 9 | 66 7 | 100 0 | 15 0 | 37 3 |
| 006 | 45 2 | 12 5 | 35 4 | 11 9 | 29.8 | 4 6 | 5 1 | 12.9 |
| 007 | 100 0 | 3.3 | 33 3 | 33 3 | 40.0 | 0 0 | 0 0 | 120 0 |
| 008 | 22 3 | 10 4 | 36 7 | 4 9 | 20 5 | 18 4 | 21.5 | 16 4 |
| 009 | 8 0 | 5 5 | 20 4 | 11 5 | 14 0 | 29 4 | 34 9 | 25 8 |
| 010 | 0 5 | 2 3 | 3 7 | 11 4 | 20 8 | 7 3 | 31 4 | 7 3 |
| 011 | 30 3 | 42 9 | 0 0 | 40 7 | 18 4 | 54 5 | 44 4 | 28 0 |
| 012 | 9 3 | 46 2 | 98 7 | 7 6 | 51.4 | 0 0 | 56 5 | 7.9 |
| 013 | 29 2 | 26 4 | 33.3 | 4 8 | 36 1 | 69 4 | 8 3 | 0 8 |
| 014 | 3 4 | 4 1 | 23 5 | 6 8 | 43.8 | 58 9 | 27 7 | 3.9 |
| 015 | 6 4 | 1 1 | 8 9 | 0.9 | 20 1 | 14 0 | 3 6 | 3.5 |
| 016 | 3.9 | 11.8 | 6 6 | 8.0 | 47.1 | 8 1 | 0 8 | 40 8 |
| 017 | 17 6 | 14 8 | 9 9 | 2 3 | 27.9 | 26 7 | 9 4 | 20 3 |
| 018 | 50 0 | 0 0 | 53.8 | 51 8 | 31 6 | 0.0 | 0 0 | 9 1 |
| 019 | 9 6 | 16 0 | 17 6 | 6 4 | 25 7 | 23 5 | 2 5 | 12.0 |
| 020 | 51 0 | 116 7 | 37 8 | 0 7 | 55 6 | 100 0 | 37.5 | 10 8 |
| 021 | 47 1 | 11 3 | 37.8 | 6 6 | 27 7 | 57.4 | 3 5 | 20 5 |
| 022 | 40 0 | 0 0 | 64 3 | 61 4 | 116 7 | 0 0 | 0.0 | 100 0 |
| 023 | 100.0 | 0 0 | 32 7 | 16 4 | 36 4 | 8.5 | 3 0 | 21 4 |
| 024 | 45 0 | 0 0 | 6 7 | 2 2 | 31 6 | 0.0 | 0 0 | 31 3 |

| External Stations - Origins | External Stations - Destinations | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 3 | 5 | 8 | 10 | 11 | 13 | 14 |
| 1 | | 37 5 | 7 4 | 0 4 | 0 4 | 20 8 | 38 9 | 0 0 |
| 3 | 95 5 | | 10 6 | 2 2 | 2 2 | 10 5 | 33 3 | 13 9 |
| 5 | 7 4 | 18 4 | | 33 3 | 0.9 | 59.7 | 15 4 | 21 1 |
| 8 | 1 1 | 6 8 | 15 9 | | 10 3 | 2 4 | 10 0 | 2 9 |
| 10 | 16 9 | 0 5 | 8 0 | 40 0 | | 43.8 | 100 0 | 6 3 |
| 11 | 27 3 | 8 7 | 100 0 | 63 2 | 0 0 | | 0 0 | 13 3 |
| 13 | 100 0 | 69 4 | 42.2 | 100 0 | 100.0 | 83 3 | | 100.0 |
| 14 | 0.0 | 0 5 | 9 3 | 11 5 | 4 4 | 72.5 | 8 3 | |

there is indication that the average sampling errors may be somewhat greater than shown in Table 1, but are well below the probability limits for average error shown there. It would follow from the theory that $\hat{K}$ for Sample 1 should turn out to be somewhat smaller than 1. It is not clear whether its failure to do so can be simply attributed to the chance fluctuation expected in $\hat{K}$, or to the uncertainty as to the true nature of sample 1 relative to volume clusters. Sample 4 produced $\hat{K}$ values of the order expected in 25 percent cluster sampling as indicated in Table 1.

## Empirical Expected Error Curves

In Equation 21 we may take the logarithm of each side to obtain

$$\log[CV(\hat{T})] = \tfrac{1}{2} \log K - \tfrac{1}{2} \log T. \qquad (26)$$

When plotted on double logarithmic coordinates, this function is a straight line with slope $-\tfrac{1}{2}$ and ordinate intercept $\sqrt{K}$. In accordance with the theory it is clear that K is a function of $\frac{1-r}{r}$, where r is the sampling rate. We write

$$\hat{K} = K^{*}\left(\frac{1-r}{r}\right) \qquad (27)$$

and substitute for K in Equation 26 to have

$$\log[CV(\hat{T})] = \tfrac{1}{2} \log K^{*}\left(\frac{1-r}{r}\right) - \tfrac{1}{2} \log T. \qquad (28)$$

Figures 4 to 7 illustrate Equation 28 for the four samples studied. The solid lines on these figures are obtained from Equation 28 when the observed average $\hat{K}$ values are used. Dotted lines on these figures represent extrapolations to other sampling rates via Equation 27 and 28. The $K^{*}$ values are also shown in Table 7.

TABLE 7

VALUES OF K AND STANDARD ERRORS

| Sample No | Universe | Actual Rate of Sampling | $\hat{K}$ | Avg $\hat{K}$ | $K^{*}$ | Standard Error of $\hat{K}$ |
|---|---|---|---|---|---|---|
| 1 | Lebanon<br>Kokomo | .499<br>500 | 1.24<br>1 21 | 1 22 | 1 22 | 0 22 |
| 2 | Lebanon<br>Kokomo | 496<br>501 | 1 21<br>1 16 | 1.18 | 1.18 | 0 21 |
| 3 | Lebanon<br>Kokomo | 483<br>491 | 1.28<br>1 41 | 1 34 | 1 28 | 0 22 |
| 4 | Lebanon<br>Kokomo | 242<br>247 | 4 59<br>3.93 | 4.26 | 1.38 | 0 92 |

### Probability Limits for Error

All of the discussion thus far has been with respect to the average sampling error to be expected. In order to give probability limits for these errors we must assume some probability distribution for the errors which can arise when a particular universe is sampled repeatedly in the same way. It is more or less conventional to suppose that 95 percent of such errors would not exceed $2CV(\hat{T})$. If the errors are normally distributed, this supposition is very close to being correct. These $2CV(\hat{T})$ probability limits are easily obtained from Figures 4 to 7 since one can simply double the $CV(\hat{T})$ corresponding to a particular sampling rate curve for a given value of T. Other multiples of $CV(\hat{T})$ can be used for different probability limits. For example, we should expect 90 percent of all possible sampling errors in a given cell to be less than $1.65\ CV(\hat{T})$, or 99 percent to be less than $2.58\ CV(\hat{T})$. The assumption of normal distribution of errors is not valid for small, and consequently unimportant, T values, say for T less than 20 trips.

When one plots the observed errors of Table 6 (relative to Table 5) on Figure 5, it turns out that about 3 percent of the 632 errors are beyond $2\ CV(\hat{T})$ limits.

## PART II PRACTICAL APPLICATION

H. L. Michael and R. M. Brown

● THE theory study resulted in a group of curves (Figs. 4, 5, 6, and 7) which are of great interest. These curves indicate, for the various methods, the average error that can be expected for any origin and destination table cell volume (that number of trips in one direction between an origin and a destination area) for various sampling rates. The data for any other sampling rate can also be obtained by interpolation. In order to use these curves for a proposed survey, two decisions must be made: (1) What minimum O-D cell volume is going to be important in this survey? (2) What average error is allowable for this minimum cell volume?

In order to answer the first of these questions, one must consider the purpose of the survey and that minimum volume of traffic which will receive important consideration because it would affect any recommendations that might be made. It is generally recognized that the low-volume traffic movements contribute very little to the overall analysis of the data or the recommendations made from that data. This does not mean that these volumes of 1, 10, 20, 40, or 50 movements are disregarded entirely, but only that they do not materially affect the survey recommendations. These small volumes will also be obtained, but they will, on the average, contain larger percentage errors.

After the minimum, important O.D. cell volume has been selected, the average error due to sampling that will be allowable must be selected. In the taking of any sample, other than a 100-percent sample, some error due to not interviewing every person will probably occur. Some O.D. cells will have less volumes than the expanded sample data indicate; other cells will have larger volumes than the sample results show. If these errors are considered as a percentage of the actual O.D. cell volume with regard only to their magnitude and not to their sign (minus percent or a plus percent) and are averaged, the average error as used in this paper will result. If the signs of the errors are considered, the average error will, of course, tend to be zero as the negative errors will tend to cancel out the positive errors.

The selection of this average allowable error will also only be for the minimum important cell volume selected. Higher volumes, according to the curves, for the same rate of sampling will have lower average errors when expressed as percentages. The percentage error allowed will again depend on the use of the survey and on the effect that such error will have on the final results. Since many O.D. cell volumes are often used together in the analysis of data, it should be noted that the errors, since some are negative errors and some are positive errors, often tend to cancel each other. Furthermore, an average error of 10 percent in a volume of 100, for example, would be only ten, a variation that might be less than the variation found if the interviewing were conducted at a different time.

It should also be noted that the total

traffic volume at an interview station has little to do with the selection of a sampling rate. The effect of the total volume is negligible if the cell frequency is small relative to the total volume. Furthermore, the sampling error in a cell which is large compared to volume is overestimated by not taking volume into account. A sampling rate, therefore, is on the safe side when it is based only on an important minimum O. D. cell volume and on an allowable average error for that volume. It is evident, therefore, that the selection of larger tracts with correspondingly larger traffic volumes would be advantageous for accuracy reasons. Here again the uses of the survey data and the accuracy needed must be considered.

The curves developed from the Lebanon and Kokomo data and shown in Figures 4, 5, 6, and 7 are, of course, only truly applicable for those two cities. There is no evidence, however, to indicate that these two cities are distinctly different in the manner of traffic characteristics from hundreds of other cities in the United States. This fact, together with the weight of mathematical theory indicated in Part I, seems to indicate that the curves as developed are applicable to surveys in other cities.

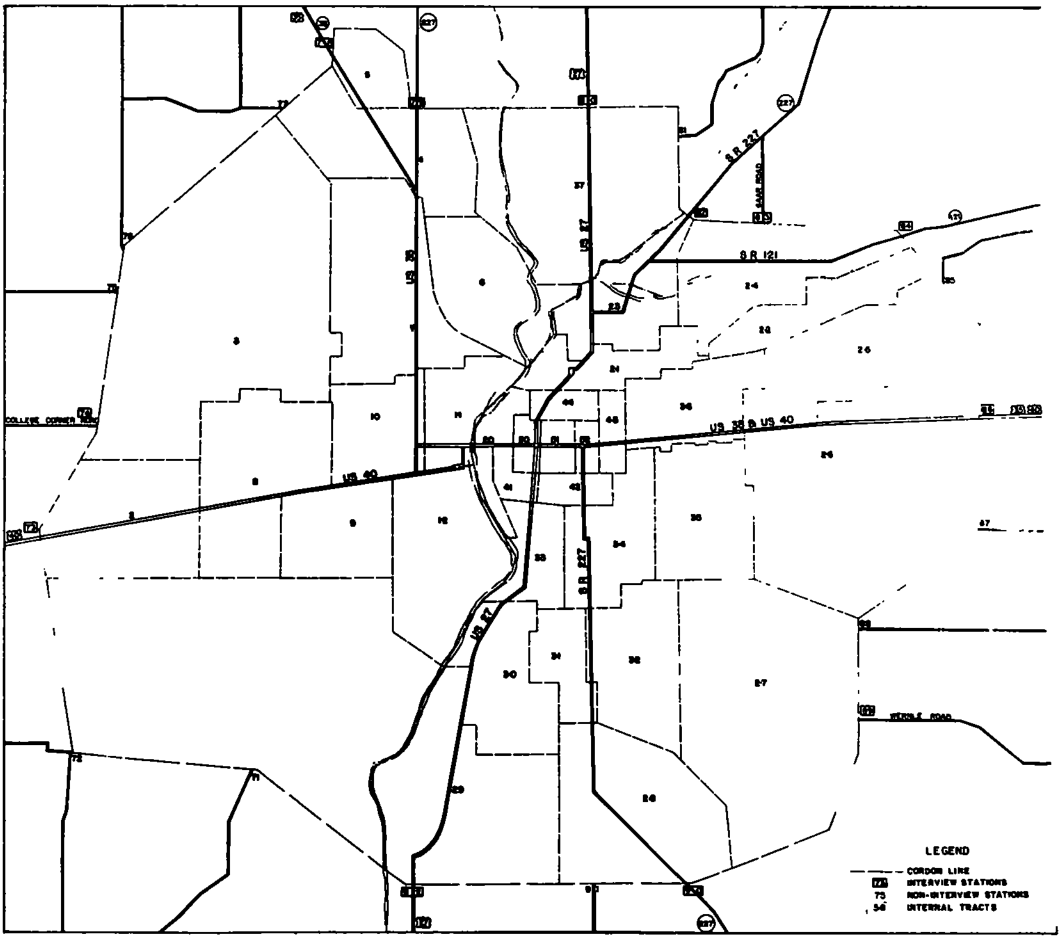


Figure 8. Richmond, Indiana, major routes and survey area.

## Use of Sampling Curves, Example

The State Highway Commission of Indiana authorized a complete origin-and-destination study for the City of Richmond, for the summer of 1952. This offered an excellent opportunity to apply the various sampling methods. The representatives of the cooperating agencies, the Metropolitan Area Traffic Survey Unit of the State High-

TABLE 8

ROAD AND TRAFFIC CONDITIONS AND SAMPLING METHOD USED
RICHMOND TRAFFIC SURVEY

| Sta No | Road Conditions: Surface | Shoulders | Land use | Average 24-Hour Traffic Volume | Inter-viewing Hours | Sampling Plan Used |
|---|---|---|---|---|---|---|
| 71 | Gravel | Narrow | Open Fields | 262 | None | None |
| 72 | Gravel | Narrow | Open Fields | 137 | None | None |
| 73 | 3-Lane Concrete | Wide | Open Fields | 11,907 | 6am - 6 am | Vol Cluster |
| 74 | Asphalt | Narrow | Open Fields | 369 | 6am -10pm | 100% Sample |
| 75 | Gravel | Narrow | Open Fields | 61 | None | None |
| 76 | Gravel | Narrow | Open Fields | 353 | None | None |
| 77 | Gravel | Narrow | Open Fields | 69 | None | None |
| 78 | Concrete | Medium | Open Fields | 5,454 | 6am - 6am | Time Cluster* |
| 79 | Asphalt | Narrow | Open Fields | 1,114 | 6am -10pm | 100% Sample |
| 80 | Asphalt | Narrow | Residential | 3,898 | 6am - 6am | Vol Cluster* |
| 81 | Gravel | Narrow | Open Fields | 181 | None | None |
| 82 | Asphalt | Narrow | Open Fields | 1,684 | 6am - 1am | Time Cluster* |
| 83 | Asphalt | Narrow | Open Fields | 460 | 6am -10pm | 100% Sample |
| 84 | Asphalt | Narrow | Open Fields | 2,162 | 6am - 1am | Time Cluster* |
| 85 | Gravel | Narrow | Open Fields | 288 | None | None |
| 86 | Divided 4-Lane Concrete | Wide | Open Fields | 13,773 | 6am - 6am | Time Cluster |
| 87 | Gravel | Narrow | Open Fields | 93 | None | None |
| 88 | Gravel | Narrow | Open Fields | 134 | None | None |
| 89 | Asphalt | Narrow | Open Fields | 552 | 6am -10pm | 100% Sample |
| 90 | Asphalt | Narrow | Open Fields | 2,671 | 6am - 1am | Vol Cluster* |
| 91 | Gravel | Narrow | Open Fields | 250 | None | None |
| 92 | Asphalt | Narrow | Open Fields | 2,586 | 6am - 1am | Vol Cluster* |

*100% sample taken during night hours of low volume count

Richmond, Indiana, June - July, 1952

way Commission of Indiana, the Bureau of Public Roads, and the Joint Highway Research Project of Purdue University, agreed to such sampling. It was further agreed that the minimum important cell volume would be considered to be 100 vehicles and that a 10-percent average error would be allowable for this cell volume. A total of 39 tracts for Richmond, a city of 38,000 people, was selected on the basis of land use and intensity of development.

A study of the curves for quarter-hour sampling (Fig. 5) and uniform sampling (Fig. 4) shows that for a cell volume of 100 and an error of 10 percent a sampling rate of slightly greater than 50 percent is necessary. A sampling rate of 50 percent was therefore chosen as one which would be practical in application. An average error of 10 percent in cell volumes of about 110 could be expected from this sampling rate.

Richmond, is a typical midwestern city. A principal east-west highway (US 40) passes through the city and four other state and federal routes pass through or terminate in the city (see Fig. 8). A cordon line around the developed area intercepted 22 roads and highways. The traffic volumes on these roads ranged from less than 100 to more than 13,000 vehicles on an average weekday (see Table 8). The roads varied from narrow, two-lane, gravel roads to modern, four-lane divided highways (see Table 8). These great differences in traffic volumes and roadway conditions permitted a study of the various sampling methods under many conditions.

The selection of the stations which were operated was made according to the usual criteria of traffic volume. A few stations on county roads were operated that had total volumes of less than 500 in order to obtain the traffic pattern for county roads. Four other stations with volumes around 2,000 vehicles per day were operated for 19 hr., instead of the usual 16 hr. because of rather large volumes of traffic between 10 pm. and 1 am. The exact location of all stations was made on the basis of the established criteria for safety.

The personnel required to operate the largest station was the determining factor for the number of interviewers hired. Each interviewer was considered capable of taking 40 interviews per hour. The number of interviewers required for each station for each hour for a 50-percent sample was thus determined (see Table 9). The number of men provided for each shift was then determined on the basis of the men needed during most hours of each shift. This method provided a rest period for usually one man at a time during the low-volume hours of the shift. Since the peak period of 3 to 6 pm. and the period 10 to 1 am. usually required several more men than other hours of these shifts, additional men were hired and assigned as necessary to the various stations during these hours (see Table 9). The necessary manual counters and crew chiefs required for the various stations were also determined. One counter and one crew chief were normally assigned to each station, except that on those

TABLE 9

PERSONNEL REQUIRED PLAN FOR A TYPICAL STATION
RICHMOND TRAFFIC SURVEY

Station No 78 — Hours of Operation 6A M to 6 A M
Sampling Rate 50% — Sampling Method Time Cluster

| Time | Pre Survey Volume Count | Men Required: Inter-viewers | Counters | Crew Chiefs | Remarks |
|---|---|---|---|---|---|
| 6-7 am | 447 | 6 | 1 | 1 | Use Following Crew |
| 7-8 am | 305 | 4 | 1 | 1 | 1 - Supervisor* |
| 8-9 am | 262 | 4 | 1 | 1 | 1 - Crew Chief |
| 9-10 am | 304 | 4 | 1 | 1 | 1 - Counter |
| 10-11 am | 408 | 5 | 1 | 1 | 4 - Interviewers |
| 11-12 N | 203 | 3 | 1 | 1 | |
| 12-1 pm | 290 | 4 | 1 | 1 | |
| 1-2 pm | 301 | 4 | 1 | 1 | |
| 2-3 pm | 327 | 4 | 1 | 1 | Use Following Crew |
| 3-4 pm | 549 | 7 | 1 | 1 | 1 - Supervisor* |
| 4-5 pm | 527 | 7 | 1 | 1 | 1 - Crew Chief |
| 5-6 pm | 403 | 5 | 1 | 1 | 1 - Counter |
| 6-7 pm | 290 | 4 | 1 | 1 | 4 - Interviewers |
| 7-8 pm | 237 | 3 | 1 | 1 | Note 2 additional |
| 8-9 pm | 186 | 3 | 1 | 1 | interviewers from |
| 9-10pm | 177 | 3 | 1 | 1 | 3 pm - 6 pm Release 1 interviewer at 8 pm |
| 10-11 pm | 160 | 2 | 1 | --- | Use Following Crew |
| 11-12 M | 148 | 2 | 1 | --- | 1 - Supervisor* |
| 12-1 am | 124 | 2 | 1 | --- | 1 - Counter |
| 1-2 am | 33 | 2 | 1 | --- | 2 - Interviewers |
| 2-3 am | 51 | 2 | 1 | --- | Note From 1 am - 5 am |
| 3-4 am | 43 | 2 | 1 | --- | take 100 % Sample |
| 4-5 am | 40 | 2 | 1 | --- | Supervisor to direct |
| 5-6 am | 120 | 2 | 1 | --- | traffic as necessary |

Operate also Station 102 on this day

*Also Supervisor of Above Station
Supervisor to assist in interviewing during under-manned hours

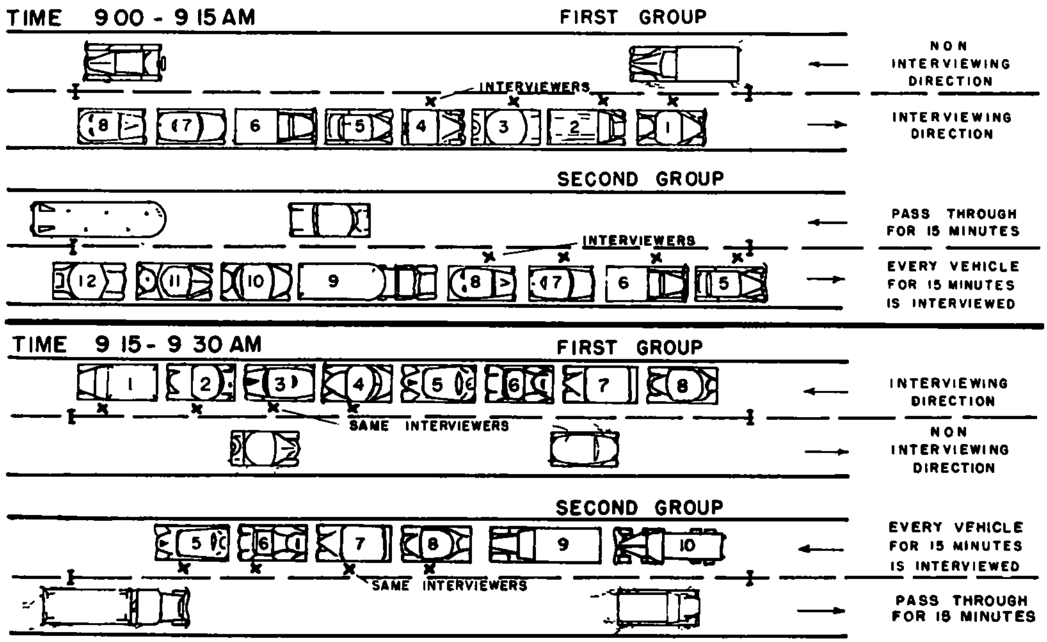


Figure 9. Typical time-cluster method of interviewing.

stations with more than 6,000 vehicles per day two counters and two crew chiefs were assigned. The crew chiefs also acted as traffic directors at heavy stations, especially during hours of peak volumes. Each shift's daily operation for all stations was also directed by a regularly employed supervisor. Typical station interview parties are also indicated in Table 9.

The preparation of such a personnel plan permitted the hiring and assignment of personnel in a most efficient and economical manner. The entire survey was conducted in seven interviewing days by a total crew for all shifts of 30 men (includes interviewers, manual counter, crew chiefs, and supervisors).

## SAMPLING PROCEDURE AT THE ROADSIDE

Three sampling methods were used during the Richmond survey: (1) quarter-hour samples (time clusters), (2) uniform samples (volume clusters), and (3) combination of (1) and (2).

The method used at each station was predetermined. Only one method of sampling was used at a station so as to minimize the confusion that might occur if sampling methods were changed hourly. Sampling was conducted, at all stations that had daily traffic volumes greater than 1,200 vehicles, during all hours that traffic volumes were expected to exceed 30 vehicles per hour in either direction. Sampling was not done at other stations or during other hours because one interviewer for each direction had to be present anyway and he could easily interview all traffic.

### Time-Cluster Sampling

The time-cluster sample as used here refers to the selection of a sample of the total traffic volume on the basis of a selected time interval. Any time interval could, of course, be chosen for this method. Every vehicle that arrives at the station during a 1-min., 5-min., 15-min., or other chosen time period is interviewed and then no vehicles are interviewed during a corresponding time period, the length of which is dependent on the selected sampling rate. If, for instance, a sampling rate of 50 percent is chosen, then interviewing is conducted for all those vehicles that arrive at the station during 50 percent of the total time. The time interval chosen for the Richmond survey was 15 min. During one 15-min. period all vehicles that arrived at the station from one direction were interviewed and no vehicles were interviewed in that direction during the next 15-min. period. This process was then repeated for all periods of the day. The interviewers, however, were not idle every other quarter hour, but interviewed traffic first for 15 min. in one direction and then for 15 min. in the other direction. A typical example of this procedure is shown in Figure 9.

Each station was arranged so that interviewing for both directions could be carried on from the same location. The station arrangement was exactly the same as if both directions would have been interviewed at the same time, except that the stop sign for the direction not being interviewed was covered during the periods when the traffic could proceed without stopping. On four-lane roads the interviewing was conducted in two lanes in each direction while on other roads, including three-lane roads, interviewing was done in one lane only for each direction.

Adequate personnel were assigned each shift, with added personnel during peak hours, to interview all traffic in the one direction. At the end of one time period in one direction and the simultaneous beginning of a time period in the other direction, interviews were of necessity taken in both directions at the same time. It is necessary in the time cluster sample to interview all traffic that arrives during the time interval. This means that at the end of some time periods there will be some vehicles that arrived before the end of the period that are waiting to be interviewed. They must be interviewed. In the other direction, there are vehicles arriving after the beginning of the next period which also must be interviewed. This problem was solved at Richmond by taking a portion of the interviewers from the direction being interviewed about 1 or 2 min. before the end of the period and placing them in position to interview the traffic from the other direction at the beginning of the next period. The other interviewers continued in the first direction until all vehicles that arrived before the end of the period had been interviewed. They then began interviewing in the other direction. This system operated smoothly at Richmond, and in every case the changeover was begun and completed in less than 5 min. (2 min. prior and 3 min. after). Usually the total delay was only a matter of 2 or 3 min.

The crew chiefs, of which there was always one and, on heavy stations, two, in addition to their normal duties of interview collection and review and general supervision of the station were given the following responsibilities: (1) direct traffic as necessary (primarily traffic not being interviewed but passed through the station and, on four-lane roads, that traffic being interviewed in more than one lane); (2) announce the end of one time period and

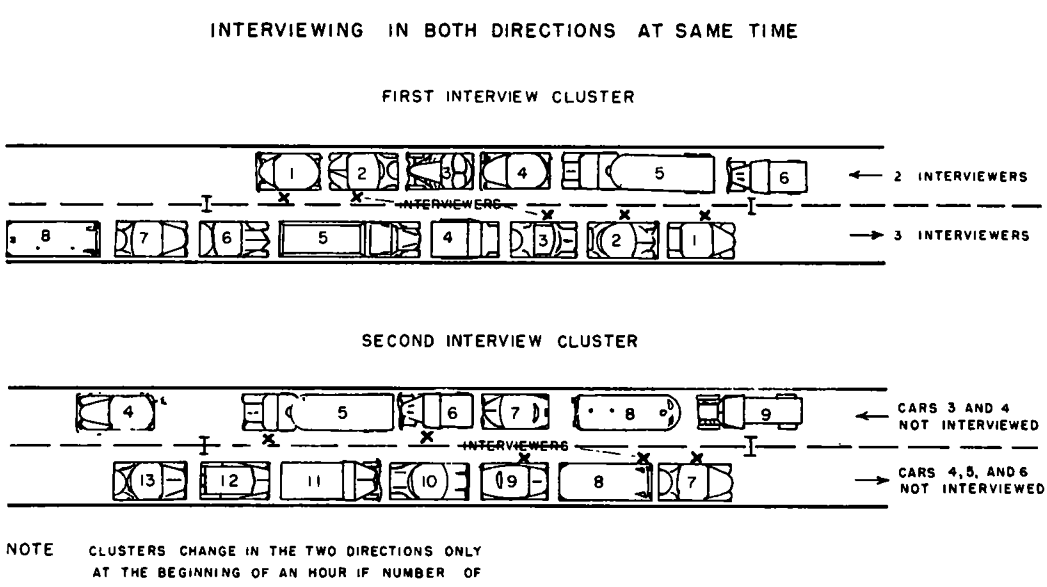


Figure 10. Typical volume-cluster method of interviewing.

the beginning of the next; (3) indicate to the interviewers the first and last car to be interviewed in each time period; and (4) select the number of specific interviewers to be taken from interviewing in the one direction before the end of the time period to get in position to begin interviewing in the other direction (number dependent on traffic volumes in both directions).

The crew chief or chiefs also had the assistance of the supervisor during peak periods and at heavy stations. All personnel were instructed to adhere to exactly 15-min. intervals.

### Volume-Cluster Sampling

The volume-cluster sampling consists of interviewing consecutive vehicles in clusters or groups of a predetermined number and then not interviewing a cluster or group of vehicles of a size that is dependent on the sampling rate. The size of the cluster to be interviewed could be one or more. A single cluster size was not used in the Richmond survey, but a cluster equal to the number of interviewers working in each direction was used. Cluster sizes or number of interviewers, however, were not changed during any one hour of operation but only at the beginning of an hour, if necessary. Cluster sizes depended on the volume of traffic expected (from previous observations and counts) in any one direction and were not the same during many hours for the two directions of one station. The actual clusters used ranged from one (every other vehicle for a 50-percent sampling rate) to six. Since all interviewing by this method was at a 50-percent rate, a cluster of vehicles equal in number to the cluster interviewed was passed through the station without being interviewed. Figure 10 illustrates this type of sampling.

The station arrangement and personnel to operate the station were the same as for time cluster sampling. Interviewers were placed in the two directions by the crew chief or chiefs as volumes in each direction warranted, except that any changes were made only at the beginning of an hour. The first interviewer in each direction and the crew chief during heavy hours determined the first vehicle in each cluster that was interviewed and directed those vehicles in the cluster that were not interviewed through the station. Each interviewer made one and only one interview for each cluster that was interviewed. This helped to maintain an accurate check of those vehicles interviewed and, at the same time, distributed the responsibility of getting accurate cluster sizes. All personnel were instructed to adhere strictly to the predetermined cluster size with no exception, except in the case of emergency vehicles.

### Combination Time-and-Volume-Cluster Sampling

In order to test a method of obtaining a 25-percent sample, a combination of the time-and-volume-cluster methods was used. This method simply obtained a sample by interviewing volume clusters within a time period. In short, interviews were taken during 15-min. intervals in one direction only and then not taken in that direction for the next 15 min. Instead of interviewing every vehicle, however, within the 15-min. period, only one half of the traffic was interviewed. This was accomplished by interviewing a cluster of traffic equal in number to the number of interviewers and then not interviewing an equal-sized cluster.

The procedure of station operation was the same as that for the time-cluster operation, with the modification that it was necessary for the first interviewer or the crew chief to pass a cluster through without interviewing and then to select the first car in each cluster to be interviewed. Only a cluster size of four was used with this method. A separate crew was used for operation in the other direction during alternate 15-min. periods. This procedure eliminated a quick change of direction for interviewers at the end of one period and the beginning of another, and gave each group of interviewers a short period of rest during each half hour.

## APPRAISAL OF SAMPLING METHODS

The success of these various methods was established after the first day for each method. The taking of a predetermined sample not only proved to be practicable but, it appeared, actually improved the interviewing process. The general public understands that there is such a thing as sampling and that it certainly should be more economical to in-

TABLE 10

ACTUAL SAMPLING RATES BY STATION
FOR PERIOD THAT SAMPLING WAS CONDUCTED
RICHMOND TRAFFIC SURVEY

50% Sampling Rate Planned

| Station | Sampling Method | Number Hours Sampled | Manual Count | Number Inter-viewed | Actual Sampling Rate |
|---|---|---|---|---|---|
| 73 | Vol. Cluster | 24 | 10669 | 5304 | 49.7 |
| 74 | 100% Sample | 16 | 290 | 290 | 100 0 |
| 78 | Time Cluster | 20 | 5091 | 2637 | 51 8 |
| 79 | 100% Sample | 16 | 891 | 888 | 99 7 |
| 80 | Vol. Cluster | 18 | 3535 | 1811 | 51 2 |
| 82 | Time Cluster | 14 | 1237 | 611 | 49 4 |
| 83 | 100% Sample | 16 | 413 | 410 | 99 3 |
| 84 | Time Cluster | 14 | 1439 | 725 | 50.4 |
| 86 | Time Cluster | 21 | 9954 | 5077 | 51 0 |
| | | 3 | 1486 | 365 | 24.6* |
| 89 | 100% Sample | 16 | 476 | 475 | 99 8 |
| 90 | Vol. Cluster | 14 | 1944 | 967 | 49.7 |
| 92 | Vol. Cluster | 14 | 1856 | 924 | 49.8 |

*25% Sampling rate planned

vestigate only a fraction of the total than to interview the whole. The public generally does not recognize the details necessary to take a sample; but from responses and discussions with motorists at Richmond, they view with approval a sampling method they can understand. The time-cluster and volume-cluster methods of sampling offered them that evidence. They could "see" the methodical, business-like selection of vehicles in 15-min. cluster or in one to six vehicle clusters and then the release of an equal-sized cluster. Time and time again the vehicle operators expressed approval of, in their words, "the business-like manner" in which the interviewing was conducted.

The selection of a fraction of the total also seemed to salve the ire of many of those residents who made many trips through the station in one day. They knew they had a gambler's chance to get through without being interviewed. They were gratified whenever they "won" and accepted the interview in a sportsmanlike manner when they "lost."

The actual selection of the sample did not prove difficult by any of the methods. Table 10 indicates the actual sample selected during those hours when sampling was actually used. Table 11 shows the hour-by-hour sampling rates for two stations. Variances in the time-cluster rates were expected because of the actual variances that occur for traffic volumes in 15-min. periods. The volume cluster, if all procedure had been exactly followed, would be almost exactly 50 percent for every station. Miscounts of the number of vehicles interviewed, or of the cluster passed through, or of the total at the station, or the failure of vehicles to stop are causes for variations. These causes, as indicated by Tables 10 and 11, were evidently controlled quite well. The use of competent crew chiefs and head interviewers did much to minimize the possibility of miscounts.

Although no consideration was given to vehicle type in the selection of the sample, approximately 50 percent of each type of vehicle was obtained. Table 12 lists the interviewing rate by vehicle type for two interview stations.

The use of the crew chief as a traffic director resulted in a speedier movement of vehicles through the station. He directed the vehicles to be interviewed to position speedily and efficiently; this left the interviewer more time to concentrate on taking the interview and properly recording it. He also kept the traffic flowing smoothly when not being interviewed. It appeared that the methods operated best when, especially during peak hours, he was the one who counted the vehicles that were not interviewed or selected the first and last vehicles that were interviewed. A crew chief as a traffic di-

TABLE 11

ACTUAL SAMPLING RATES BY HOUR AT TYPICAL STATIONS
FOR RICHMOND TRAFFIC SURVEY

| Station No | 73 | | | 86 | | |
|---|---|---|---|---|---|---|
| Sampling Method | 50% Volume Cluster | | | 50% Time Cluster | | |
| Hour | Manual Count | Number Inter-viewed | Actual Sampling Rate | Manual Count | Number Inter-viewed | Actual Sampling Rate |
| 6-7 am | 522 | 264 | 50 6 | 511 | 284 | 55 6 |
| 7-8 am | 507 | 254 | 50 1 | 454 | 229 | 50 4 |
| 8-9 am | 529 | 265 | 50 1 | 507 | 255 | 50 3 |
| 9-10 am | 558 | 279 | 50 0 | 492 | 269 | 54 7 |
| 10-11 am | 554 | 280 | 50 5 | 540 | 273 | 50 6 |
| 11-12 N | 545 | 266 | 48 8 | 600 | 300 | 50 0 |
| 12-1 pm | 561 | 283 | 50 4 | 582 | 287 | 49 3 |
| 1-2 pm | 548 | 264 | 48 2 | 616 | 306 | 49 7 |
| 2-3 pm | 565 | 275 | 48 7 | 598 | 321 | 53 7 |
| 3-4 pm | 663 | 331 | 49 9 | 684 | 366 | 53 5 |
| 4-5 pm | 776 | 392 | 50 5 | 767 | 398 | 51 9 |
| 5-6 pm | 706 | 351 | 49 7 | 717 | 372 | 51 9 |
| 6-7 pm | 559 | 270 | 48 3 | 632 | 290 | 45 9 |
| 7-8 pm | 605 | 299 | 49 4 | 594 | 294 | 49 5 |
| 8-9 pm | 550 | 283 | 51 5 | 668 | 322 | 48 2 |
| 9-10 pm | 446 | 216 | 48 4 | 574 | 134 | 23 3* |
| 10-11 pm | 365 | 181 | 49 6 | 492 | 124 | 25 2* |
| 11-12 M | 290 | 142 | 49 0 | 420 | 107 | 25 5* |
| 12-1 am | 200 | 99 | 49 5 | 242 | 128 | 52 9 |
| 1-2 am | 131 | 65 | 49 6 | 189 | 101 | 53 4 |
| 2-3 am | 106 | 50 | 47 2 | 114 | 54 | 47 4 |
| 3-4 am | 121 | 61 | 50 4 | 105 | 61 | 58 1 |
| 4-5 am | 110 | 58 | 52 7 | 122 | 68 | 55 7 |
| 5-6 am | 152 | 76 | 50 0 | 220 | 99 | 45 0 |

*25% Sampling rate planned

rector appeared to be a must. Most of his other duties, such as collection and review of the interviews, were handled by the shift supervisor on heavy volume stations.

The use of a predetermined sample also minimized the confusion at interview stations. Each man knew what he was to do, who he was to interview, and who he was not to interview. During peak hours there was no doubt or confusion as to how long the line of waiting traffic should get before releasing some or how many to pass through without interviewing, since the interviewers were instructed to take the predetermined sample according to the predetermined method at all times. During peak hours some back-up of traffic did occur. It appeared, however, that a line of traffic formed almost instantly (such as would occur from a large group arriving simultaneously), then stayed about the same length for a short period, and gradually disappeared within 5 to 10 min. This cycle, during peak periods, then often repeated itself. The maximum delay for any vehicle under either method was slightly less than 5 min. This maximum delay occurred only during very short periods of time during peak hours on high volume stations (12,000 vehicles per day). Many times no delay except for interviewing occurred, and during most hours of the day the maximum delay was two to 3 min.

TABLE 12

SAMPLING RATES FOR TYPICAL STATIONS BY TYPE OF VEHICLE FOR RICHMOND TRAFFIC SURVEY

| Station No | 73 | | | 86 | | |
|---|---|---|---|---|---|---|
| Sampling Method | 50% Volume Cluster | | | 50% Time Cluster | | |
| No. of Hours Sampled | 24 | | | 21 | | |
| Vehicle Type | Manual Count | Number Interviewed | Sampling Rate | Manual Count | Number Interviewed | Sampling Rate |
| Passenger Cars | 8455 | 4199 | 49 7 | 7808 | 3975 | 50 9 |
| Light Trucks | 700 | 340 | 48 6 | 465 | 247 | 53 1 |
| Medium Trucks | 480 | 258 | 53 8 | 514 | 263 | 51 2 |
| Heavy Trucks | 970 | 479 | 49 4 | 1104 | 560 | 50 7 |
| Buses | 64 | 28 | 43 8 | 63 | 32 | 50 8 |

The volume-cluster method appeared to work most efficiently and smoothly on two-lane roads where four interviewers or less in each lane handled the volume of traffic. More interviewers than this resulted in a decrease in the number of interviews an interviewer took per hour because of the longer lapse between interviews, due to passing a large cluster through the station. The interviewing efficiency did not become serious, however, unless more than six interviewers per direction were used, since the traffic director speeded the movement through the station. This method worked better than the time cluster on narrow roads with poor shoulders, because all interviewers were permanently positioned and required less necessary space in which to operate. On the three-lane road this method, with interviewing in one lane per direction, worked quite well with as many as six interviewers on line in one direction.

The time-cluster method also worked best where the number of interviewers was four or less per interviewing lane. With this method, since all vehicles were taken for a selected time period, speed of interviewing was most important in order to prevent long lines of waiting traffic. This method required a little more working space for the interviewers, since they were required to change position every time period and since one direction of traffic was moving freely through the station. A little spare space in the center of the highway was advisable for obvious safety reasons. On the four-lane highway, with two interviewing lanes for each direction, this method worked excellently. A short time before the end of the time period, the inside lane was closed to interviewing and the interviewers on that lane got into position to interview in the outside lane of the other direction. After all required vehicles were interviewed in the first direction, the remainder of the interviewers there transferred to the inside lane of the second direction. Traffic in the direction that was not interviewed passed through the station in the outside lane. This made it unnecessary for interviewers to cross lanes of moving traffic. A careful plan of action for the short period at the end of one time period and the beginning of the next was extremely important for the time-cluster method in order to minimize confusion and to select all the proper vehicles in the sample.

A time period of any length could, of course, be chosen. The 15-min. period was selected for this survey for at least two reasons:

1. The original study of the Lebanon and Kokomo data was made for 15-min.

time clusters, since the field data were separated only for each quarter hour. This time period, incidentally, proved to be the most accurate of all the methods studied with this data.

2. A shorter time period would have necessitated more frequent directional changes of interviewing. This has definite disadvantages as it increases chance of confusion and also has the statistical objection of increasing the number of times that the selection of the first and last vehicle is left up to an individual who may err as to its proper selection.

The selection of a shorter time period, however, might reduce the length of any waiting traffic lines. It appears that several studies of a shorter time cluster should be made.

It must be admitted that time- (and possibly volume-) cluster samples would have biases if a succession of cars would all have the same origin-destination. If such movements were more likely to happen at a certain time than at another, a bias might occur if interviewing always began at a certain hour on the hour. One way to try to minimize such biases would be to leave to chance (such as spin a needle around the clock) when the interviewing should begin. However, biases eliminated in some origin-destinations might be put into others in this way. It appears, therefore, that on the average starting at the beginning of an hour would be as good as could be done. From a practical point of view, it certainly would be best.

The combination time-and-volume-cluster sample, although only used for 3 hr., appeared to operate smoothly and efficiently. A 25-percent sample was taken without delaying traffic for long periods and with a small group of interviewers who obtained an average number of interviews per hour. Under the volume-cluster plan, the number of interviews per interviewer for this rate would probably be less because of the necessarily long waits between interviews while large clusters of traffic not being interviewed passed through the station. On the other hand, the use of the time-cluster method only would necessitate periods of time when the interviewers would not work at all. The use of the two methods together minimizes these disadvantages, but at the same time makes a changeover to another direction at the end of a time period difficult. A solution to this problem might be to allow a short time lapse between periods when no interviews for either direction would be taken. This, of course, would reduce the sampling rate and would have to be considered in obtaining the desired size sample.

It must be emphasized that thorough planning is necessary to insure proper operation of any of these sampling plans. The required number of interviewers must be determined from presurvey counts and then that number selected and thoroughly instructed. Competent crew chiefs and head interviewers are also a necessity. Above all, every one connected with the operation must realize the importance and understand the selection of the sample.

## CONCLUSIONS

1. If the traffic passing through a roadside interview survey station is sampled in clusters, on either a time or volume basis, then, for the numerous estimates of O.D. trip frequencies associated with this traffic flow, the overall sampling variation is similar to that given by the theory of random sampling. When a delineation is made among the different types of cluster samples, experimental results from two independent surveys are in general accord with the corresponding sampling theory derived from basic assumptions on the nature of the sampled universes.

2. The expected percentage sampling error, $CV(\hat{T})$, for an estimate, $\hat{T}$, of a true O.D. frequency, $T$, can be expressed by the relation

$$CV(\hat{T}) = (K/T)^{1/2}$$

where $K$ depends primarily upon the sampling rate, and, to a lesser degree, upon the method of sampling. Empirical values of $K$ were obtained in this study and may be used to estimate the magnitude of sampling errors and their limits in probability. Conversely, the results can be used to select a sampling plan when a specified amount of sampling variation is to be allowed for a particular origin-destination survey.

3. It is evident that the use of a predetermined sampling process is both practicable and beneficial. A 50-percent sample was obtained with a minimum of confusion in a business-like manner that was recognized and appreciated by the partici-

pating vehicle operators and the community.

4. The sampling procedure promoted efficient operation of the interviewing station by minimizing confusion, and assured the selection of a sample whose chance variation could be appraised.

5. The time-cluster method performed more smoothly than the volume method on roads where traffic volumes could be handled by three or four interviewers per interviewing lane and where adequate space for safe movement of interviewers between interview lanes was available.

6. The volume-cluster method operated more efficiently than the time cluster method on roads where traffic volumes required five or six interviewers or where roads and shoulders were narrow.

7. Either method operated well where only one or two interviewers were required.

8. The combination time-and-volume-cluster method appears to have possibilities where small samples (25 percent or less) are adequate.

9. A great deal of thought and study is necessary to determine the sampling accuracy which is necessary for origin-destination studies. Specifically, the question that requires study is: What is the allowable average error (or a certain probability error) for various origin-and-destination, rectangular-table, cell volumes? It may be that an average error of 10 percent for cell volumes of 100 may be more accurate than generally necessary.

10. Thorough planning of station operation and the assignment of adequate and competent personnel are required. This should include presurvey hourly traffic counts and excellent instruction of personnel.

11. The use of some sampling procedure not only will reduce the number of personnel required to conduct the field work but will also reduce the number of cards to be coded, key punched, and tabulated. Such reduction probably will be in direct proportion to the number of interviews taken. This not only reduces the cost of O. D. surveys but reduces the time necessary to complete such a study.

12. A number of other studies on similar or other sampling methods as well as studies on necessary size of sample would prove of benefit to the conduct of future origin-and-destination surveys.

## ACKNOWLEDGEMENTS

The material presented in this paper was made possible by the cooperation of many individuals. The authors are especially indebted to K. B. Woods and A. K. Branham of the Joint Highway Research Project, Purdue University, for their constant encouragement and helpful guidance: to A. D. May, Jr., formerly of the Joint Highway Research Project, for his interest and assistance in the initial phases of the study; and to S. F. Coffman, Bureau of Public Roads, for his invaluable advice and suggestions during the practical application study. Special acknowledgement is also given to many departments of the State Highway Commission of Indiana for their assistance and cooperation. To the many individuals who assisted with the preparation of the data and the report, the authors also wish to express their personal gratitude.

## REFERENCES

1. Deming, W. E., Some Theory of Sampling, J. Wiley and Sons Inc., New York (1950).

2. Kennery, J. F. and Keeping, E. S., Mathematics of Statistics, Part Two, D. Van Nostrand Co., Inc. (1951).

3. Michael, H. L., "Traffic and Engineering Report of the Before and After Surveys of the Kokomo By-Pass", Thesis submitted to Purdue University in partial fulfillment of the requirements for the degree of Master of Science in Civil Engineering (1951).

4. Miller, I., "The Application of Sampling Techniques to Highway Surveys", Thesis submitted to Purdue University in partial fulfillment of the requirements for the degree of Master of Science (1952).

5. Manual of Procedures for the Metropolitan Area Traffic Studies, Public Roads Administration (1946).

6. Manual of Traffic Engineering Studies, Association of Casualty and Surety Co. (1945).