THE SAMPLE DESIGN FOR THE 1957-58 DETROIT AREA STUDY

The 1957-58 Detroit Area Study used a relatively new sample design which resulted in a close approximation to a simple random sample of Detroit area residents. The computation of sampling errors and the application of statistical techniques to the data which came out of this sample is therefore a much simpler process than when a more clustered sample is employed.

Description of Sample Area

The sample area comprised the sections of Macomb, Oakland and Wayne counties that are divided into census tracts by the U. S. Bureau of the Census. The area covered represents approximately 87 per cent of the total population of the three-county Detroit Standard Metropolitan Area. The nontracted area was not included in the sample because of the higher cost of securing interviews in this area relative to areas closer to Detroit, and because of the small number of interviews which would have been obtained compared to the number obtained in the tracted area.

Determination of Sample Size and Sampling Rate

Budgetary factors and considerations of desired precision led to the decision to aim for an interview return of 700. In designing the sample to yield 700 interviews, two factors were given careful consideration: (1) the desired response rate at occupied sample addresses, and (2) the rate which all addresses in the universe can actually be found through our sampling techniques (the coverage rate). The response rate for our 1957-58 survey was set at 85 per cent (i.e., out of every one hundred eligible respondents who fell into the sample, we expected to interview eighty-five). The coverage rate was estimated
at 97 per cent (i.e., out of every one hundred dwellings in the universe, we expected that ninety-seven would be represented in the sample). Hence the fraction, \( \frac{700}{.97} = 821 \), indicates that 821 sample addresses would be needed to yield 700 completed interviews, given a 100 per cent coverage. It was estimated that 936,791 occupied dwelling units were in the research universe as of February 1, 1958. Thus, 936,791 multiplied by .97 (estimated coverage rate) equals 908,687 dwellings represented by our sampling techniques. The overall sampling rate of one in 1103 (which was rounded down to 1100) was determined, therefore, with this fraction: \( \frac{936,791 \cdot .97}{700 \cdot .85} = \frac{908,687}{821} = \frac{1103}{1} \). This sample was designed to provide a master sample for three years; the actual rate of sampling was therefore 3 in 1100.

**Delineation of Primary Sampling Units**

The Detroit area was divided into primary sampling units (PSU's) so that the area to be sampled was exhausted and each PSU was unambiguously defined. PSU's were comprised of incorporated cities and tracts. The delineation was made on the basis of boundaries established by the 1950 Census.

**Stratification of Primary Sampling Units**

In this sample there were two types of primary sampling units, those which were self-representing and those which were non-self-representing in that they represented a larger area. Self-representing PSU's consisted of all incorporated cities with a 1950 population of 25,000 or more; the non-self-representing PSU's were selected from strata composed of cities of smaller population and unincorporated tracted areas. In this way, some stratification according to population was accomplished.
For the area of the sample not included within self-representing PSU's, further stratification on the basis of geographic and economic categories was utilized to reduce sampling error while maintaining the representativeness of the sample. The tracts and small cities which composed the remaining PSU's were given economic ratings of high, medium, or low based upon a combined measure of median family income and rental and house value. PSU's in similar geographic locations (e.g., those lying north of the City of Detroit, etc.) were assigned the same geographic rating. PSU's with similar geographic and economic ratings were then grouped in the same stratum.

**Estimation of PSU Population Size**

For probability sampling, a reasonably accurate estimate of each PSU's population or occupied dwelling units is required to assign probabilities and to minimize sampling variations. Such estimates were available from the Detroit Metropolitan Regional Planning Commission. For most of the PSU's, however, the data available were for July 1, 1956 and 1957 while the desired estimate was needed for February 1, 1958, the date of field interviewing.

To estimate the ODU's in an area, we assumed an increase between July 1, 1957 and July 1, 1958 of the same percentage magnitude as the increase from July 1, 1956 to July 1, 1957. Furthermore this increase was assumed to be linear throughout the year. On the basis of these assumptions, \( \frac{7}{12} \) of the estimated ODU increase of an area for the period July 1, 1957 to July 1, 1958 would give an estimate for an area for February 1, 1958. These figures were checked with similar projections made by the Sampling Section of the Survey Research Center and found to be in substantial agreement.

For some tracts, precincts, and enumeration districts the data were given only for a larger area. To obtain estimates for these smaller divisions.

*These ratings are those used in the Detroit area by the Sampling Section of the Survey Research Center and, in turn, derived from 1950 Census data.*
various techniques were used; the major procedure being to assume that the sub area would contain the same percentage of the ODU's as they did in the 1950 Census. The 1958 estimate of the ODU's was distributed accordingly.

Selection of the PSU's to Represent Stratum

The selection of the primary sampling units from all strata was accomplished in the following manner. Each PSU in a stratum had an ODU population estimate as of February 1, 1958. A cumulative total for each stratum was computed from these estimates. Consulting a table of random numbers, the first table number between one and the cumulative total for the stratum selected the PSU, which contained the random figure in the cumulative column, to represent a specific stratum.

Strata vary in size, as do the PSU's selected to represent these strata. The probabilities of selection of the dwelling units, however, should be the same. In order to accomplish this, the probabilities of the various steps in the selection process must equal the overall sampling rate \( \frac{3}{1100} \). As will become clear later, the selection of sample PSU's is but one stage in the sampling equation and represents the first probability to be computed. Later stages of the sampling equation will be adjusted to preceding stage probabilities in order to maintain the overall sampling rate. The sampling equation at this stage was: \( \frac{3}{1100} = \left( \frac{p_i}{s_i} \right) \left( \frac{1}{x} \right) \), where the probability of the PSU being selected from a particular stratum is expressed by \( \frac{p_i}{s_i} \); \( p_i \) is the estimated ODU's in the PSU, and \( s_i \) is the number of ODU's in its stratum. The probability of selection of the remaining stages is represented by \( \frac{1}{x} \); for example, if the probability of selecting a PSU is \( \frac{1}{11} \), \( \frac{1}{x} \) will equal \( \frac{3}{100} \).
Selection of Sample Dwelling Units

Once a sample of primary units was chosen, the method of selecting the dwelling units varied with the type of selection procedure to be used. For this design two different methods of sampling were used: a City Directory sample and a segment sample.

City Directory Sample

Because of the close approximation City Directory sampling can have to simple random sampling, the City Directory method was preferred over the segment sample and was used wherever possible.

In the past, the utilization of a City Directory in selecting a sample has frequently failed to provide an adequate coverage of a population. The coverage problem was largely eliminated in the Detroit Area Study sample because we sampled addresses rather than names (much more stability being associated with addresses than with names).

Another objection has been that not all dwelling units are included in a Directory; this problem is especially important when older Directories which do not include new construction must be used. To minimize this defect, a close working relationship was established with the Polk City Directory Company in Detroit. The Polk Company kindly allowed us to use a City Directory for Detroit East Side which was only then in page proof. It may be noted that in the 1957-58 study, no Directory published before 1956 was used.

The accuracy and completeness of the City Directory listings were measured by two procedures: (1) the exhaustiveness and exclusiveness of area coverage were investigated, and (2) a block-supplement sample checked the completeness of the address coverage and brought the sample up to date. These devices are described in the next section.
Check on Area Coverage

The boundaries listed in the Directory vary from the very precise to the extremely nebulous. To insure that a complete area was covered, we compared an estimate of the number of units included in the City Directory with an estimate of the number of units expected in the area. We selected a small sample of pages from the Directory, counting the dwelling units (du's) on each page, computing the average per page and multiplying by the number of pages in the Directory; this number was then compared with population projections. We also checked certain streets to find whether or not the listing extends to the boundaries of the area covered by the City Directory. If a large discrepancy occurred between the estimates of the first method, or if the listings do not cover the street area, the City Directory will not yield a proper sample of the area.

The Directory may also include addresses outside of the boundaries of the sample area. Thus sample addresses selected from a City Directory must be checked to make sure they are within the boundaries of the sampling area. If an address falls outside the area, it must be dropped because it belongs to another PSU or is not in the overall sampling universe (i.e., the tracted area of the Detroit Metropolitan Area). The usual method of checking an address was to locate it in a street index that lists the tracts of the addresses. Because PSU's boundaries frequently follow tract boundaries, this method helped to determine if the address was in the desired area.

A further problem in City Directory sampling arises from the practice of publishing a large Directory that does not distinguish between many small incorporated and unincorporated units. This problem is especially acute where there may be two or three areas to be sampled out of six or seven covered by the Directory, and where these two or three areas represent a small percentage
of the coverage of the Directory. In such cases, a segment sample must be utilized. However, a City Directory sample may be selected if a large percentage of the total Directory covers the areas to be sampled.

**Block Supplement**

As stated above, a City Directory sample must be augmented through the use of a "block supplement." In this way, dwellings which were not listed in the Directory are given the opportunity of coming into the sample.

The first step was to list all of the blocks covered by the Detroit City Directories in a geographical serpentine order. The listed blocks were then sampled at a rate of $\frac{3}{1100}$ to provide a master sample of blocks to be used for a period of three years. The blocks were sampled at a rate of $\frac{1}{1100}$ and clusters of three (the one selected plus the next two) were taken. Since each block, therefore, had three chances of being selected in a cluster, the probability of selection was raised to the desired $\frac{3}{1100}$ for each block. (The effect of clustering on probabilities of selection, with reference to the selection of clusters of addresses from the City Directory will be discussed later.)

In the 1957-58 study, a subsample of blocks was obtained by taking every third block of the master sample listing. A City Directory listing of all the addresses found on these blocks was then prepared. Each of these blocks was then visited and every address actually found on these blocks was compared with the City Directory list. All addresses found on a given block which were not listed in the City Directory were selected for interviewing.

This procedure provides an indication of the adequacy of the City Directory and also provides a sample of the addresses missed by the City Directory.

In addition to Detroit, Dearborn was selected into the block supplement with certainty; the remainder of the PSU's sampled with City Directories were
selected by a random procedure based upon occupied dwelling unit estimates for each PSU. This resulted in Roseville and Lincoln Park being selected to represent these remaining PSU's. Within Dearborn, Roseville, and Lincoln Park, a sample of blocks was selected in the same manner as in Detroit. However no clustering was utilized which required the selection of blocks at a rate of \( \frac{3}{1100} \).

**Selection of Addresses from City Directory**

After we determined that the City Directory provided an adequate coverage, we proceeded with the actual selection of addresses from the Directory. It will be remembered that each PSU in our sample universe had a certain probability of entering the sample. In most cases each of the City Directories for greater Detroit covered a single PSU which was self-representing.

According to the sampling equation \( \frac{3}{1100} = \left( \frac{p_i}{s_i} \right) \left( \frac{1}{x} \right) \), the probability of selecting a given City Directory, if it corresponds to a PSU, is represented by \( \frac{p_i}{s_i} \). Consequently, \( \frac{1}{x} \) must be the probability of selecting addresses from the City Directory.

To clarify the process, let us use an actual sample, that of Dearborn. Being a self-representing PSU, \( \frac{p_i}{s_i} \) becomes 1. Therefore, the equation now reads \( \frac{3}{1100} = \left( \frac{1}{1} \right) \left( \frac{1}{x} \right) \) or \( \frac{3}{1100} = \frac{1}{x} \). In practical terms, this means that we want to select addresses at the rate of \( \frac{3}{1100} \). In doing this, we are selecting a master sample which may be used in subsequent years. Since the sample size for the Detroit Area Study varies little from year to year, ranging between 650 and 750 completed interviews, it was estimated that a sampling rate of \( \frac{1}{1100} \) would be required for one year, thereby requiring a rate of \( \frac{3}{1100} \) for the master sample designed to last for three years.
It would have been possible to take a systematic random sample by counting the number of addresses on each page, totalling these cumulatively, and assigning a random start and the proper interval. But this would have involved an undue amount of clerical time and energy.

An alternative procedure was used. Instead of totalling the number of addresses on each page, a sample of pages was selected at an interval which would minimize the clustering of sample addresses on a page. Also, instead of counting addresses and sampling them, the procedure was to sample spaces in the Directory. These spaces correspond in size to the space required for the listing of an address line in the City Directory. The selection of addresses is therefore a two-stage process: (1) the sampling of pages, and (2) the sampling of spaces on these pages.

Selection of Pages

Instead of selecting individual lines from the City Directory, we selected lines in groups of three (hereafter referred to as clusters), since the master sample was to be utilized in three different samples, in three successive years. Once we have selected clusters it is very easy to select one of the three lines for each of the years.

As we have said, the population estimate and desired sample size for the 1957-58 study dictated a sampling rate of \( \frac{1}{1100} \). Therefore, if exactly one-third of the lines are to be used each year with approximately the same size sample, it is necessary to select the master sample clusters at a rate of \( \frac{3}{1100} \).

The selection of clusters was accomplished by selecting individual lines at a rate of \( \frac{1}{1100} \). Having selected these lines as the first lines of each cluster, the following two lines are automatically included, thus forming clusters of three. It can readily be seen that in this procedure every line has a probability of \( \frac{3}{1100} \) of being selected. A line may fall into the sample
in three ways: (1) by being selected as the first line of the cluster, (2) by being included as the second line of the cluster, or (3) by being included as the third line of the cluster. An example will show what occurs:

1665 Wilshire Avenue
1668 Wilshire Avenue
1771 Wilshire Avenue
1811 Wilshire Avenue
1955 Wilshire Avenue

Take, for example, the address: 1771 Wilshire Avenue. It will be selected if 1665 Wilshire Avenue is selected, in which case it is the third line of the cluster. If 1668 Wilshire Avenue is selected, it will be the second line of the cluster. It also has a \( \frac{1}{1100} \) probability of being selected as the first line of the cluster.

Therefore, two facts must be held firmly in mind: (1) every cluster of three lines has a selection probability of \( \frac{3}{1100} \), and (2) this gives each of the three lines of a cluster a probability of \( \frac{3}{1100} \) of being selected.

Once the master sample of clusters had been drawn, one-third of the lines was selected for use in the 1957-58 research. In this subsampling procedure every line had a \( \frac{1}{3} \) probability of being selected from the Master Sample. This \( \frac{1}{3} \) probability when multiplied by the original line probability of \( \frac{3}{1100} \) gives each line a probability of \( \frac{1}{1100} \) of being selected in the year's sample.

For the discussion which follows it is important to remember that every cluster must be selected with a \( \frac{3}{1100} \) probability.

The process of page sampling will be described with the aid of an example from the Dearborn Directory. In Dearborn, there was an estimated 33,500 occupied dwelling units at the time of the 1957-58 study. Taking \( \frac{1}{1100} \) times this total gave us the number (31) of clusters of three lines each that must be selected from the Dearborn Directory for a three year sample.
To provide as much randomness as possible, we attempted to select pages so that there was no more than one cluster per page. This cannot be guaranteed because of the varying amount of print space on a page. But it can be approximated by dividing the number of pages in the Directory (258) by the number of clusters desired (31) to give the page interval. This resulted in a page interval of \( \frac{1}{8.3} \) for Dearborn which is the sampling rate for pages for this PSU Directory.

A random number between 0 and 8.3 provided the random start which gave the first page in the Dearborn sample. The next page to be selected was 8.3 pages from the random start and so on. By this procedure, we selected 31 pages in the Dearborn City Directory.

**Sampling Addresses**

The next stage of the 1957-58 sample required the selection of clusters of addresses. In doing this we treated the addresses found on the selected pages as one continuous list. Each of these addresses had a certain probability of being selected into the sample by virtue of their having been listed on one of the sample pages. This probability was equal to the probability of any one of the pages being selected. In our example, each of these pages had a selection probability of \( \frac{1}{8.3} \). But according to our sampling equation every cluster must have a selection probability of \( \frac{3}{1100} \). The \( \frac{1}{x} \) factor now is the probability of a cluster being drawn into the sample, so that equation now reads:

\[
\frac{3}{1100} = \left( \frac{1}{8.3} \right) \left( \frac{1}{x} \right).
\]

By solving for \( \frac{1}{x} \) we find that every cluster must have a \( \frac{3}{133} \) probability of selection from the continuous list of addresses formed by the 31 pages. Every 133rd line was therefore taken as the first line of every cluster.

This completes the process of drawing a master sample from a self-representing PSU which is adequately covered by a City Directory. The entire equation for Dearborn now reads:

\[
\frac{3}{1100} = \left( \frac{1}{8.3} \right) \left( \frac{3}{133} \right); \text{ where } \frac{3}{1100}
is the overall sampling fraction; 1 is the probability of Dearborn being
selected into the sample, \( \frac{1}{6} \) is the probability of a page being
selected from the Dearborn City Directory, and \( \frac{3}{133} \) is the probability of a
cluster being selected from all addresses on the 31 pages selected from the
City Directory. Thus, \( \frac{3}{1100} \) represents the probability of a cluster being
selected from all the addresses in the City Directory.

For non-self-representing PSU's, those which represent a stratum, the
equation is similar except that the probability of selecting that PSU will be
less than 1. It is a simple matter to adjust the other factors in the equation
to satisfy the overall sampling equation.

It should be pointed out at this time, as an example of the precision
adhered to in the sampling procedures, that the number of clusters required
from a PSU is not determined by simply taking \( \frac{1}{1100} \) times the number of
occupied dwelling units in the PSU and rounding off any fraction that might
occur to the nearest whole number. Nor is a separate random start used for
each separate PSU. If these procedures were followed, it could easily result
in constant over-sampling or under-sampling of the several PSU's, for it cannot
be assumed that either the rounding process or the separate random starts
procedure will average out in the long run.

To eliminate the possibility of over-sampling or under-sampling, a con­
tinuous list of all the selected PSU's in the sample is formed and the total
number of expected clusters ascertained for each. The total of expected
clusters was cumulated, including the fractional parts, a random start between
0 and 1 was selected and an interval of 1 was used. By this procedure, as many
clusters will be selected from a PSU as there are whole numbers included in
its range. It can be seen that the fractions will accumulate and the PSU's
which get the extra clusters (equal in number to the sum of the fractional
parts) will be determined partly by this cumulation and partly by the random start. Following is an example using a random start of 0.2 and an interval of 1.0.

<table>
<thead>
<tr>
<th>(1) PSU</th>
<th>(2) Estimated ODU's</th>
<th>(3) $Co. x \frac{1}{1000}$</th>
<th>(4) Cum. Total</th>
<th>(5) Assigned No. of Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birmingham</td>
<td>7,360</td>
<td>6.7</td>
<td>6.7</td>
<td>7</td>
</tr>
<tr>
<td>Royal Oak</td>
<td>21,490</td>
<td>19.5</td>
<td>26.2</td>
<td>20</td>
</tr>
<tr>
<td>Oak Park</td>
<td>9,046</td>
<td>8.2</td>
<td>34.4</td>
<td>8</td>
</tr>
</tbody>
</table>

In Birmingham, seven clusters were needed because the following seven values are found between 0 and 6.7 (the cumulative total for Birmingham): 0.2, 1.2, 2.2, 3.2, 4.2, 5.2, 6.2. Twenty clusters are required from Royal Oak because twenty values fall between 6.7 and 26.2.

By following this procedure we may get either 34 or 35 clusters depending upon the random start. This is a total variation of one compared to the variation of 3 which would have been possible by assigning a random start to each of the PSU's. The reader may select random starts ranging from 0 to 1 to satisfy himself on this point.

**Listing of Addresses**

To summarize the development of our City Directory sample to this point, we have determined the number of clusters required for each PSU, and we have selected pages and sample lines on those pages. We should point out again that the selected lines of the City Directories are not addresses but spaces. A space is the print space required for one address. This will vary from directory to directory and necessitates the construction of a new address space scale for each directory.

An address scale was divided into address spaces, the number of spaces being determined by the maximum length of a column of address print on a
Directory page. In other words, the advertising at top and bottom is not included as part of the printed spaces to be sampled.

The address scale was then used to determine the number of spaces in each column on each of the sampled pages of a directory. The spaces per column are totalled cumulatively; thus, every space on the sampled pages in a single Directory is given a line number running from 1 to n.

For example, let us take the first selected page of a typical directory. Let us imagine that in the first column there is room for 70 spaces. In the second column there are 70 spaces also. This makes a cumulative total of 140, so the last space in column 2 of this Directory will correspond to line 140.

If the sample had called for line 80, the column within which this line falls is located and the cumulative total of the previous column is subtracted, in this case, giving us line 10 in column 2 to be selected. The address space scale is applied to determine which address falls into the 10th space. This becomes the first address in the cluster. The addresses in the next two spaces are automatically included to provide the cluster of three addresses. If no address is found within either of the three spaces, no addresses will be selected. Nevertheless, one cluster of blank directory spaces will have been obtained.

One minor clerical problem results from the fact that since the addresses are not evenly spaced, it frequently occurs that a space may include only part of an address. To account for this without biasing the sample the following rule was followed: Any address which has any part included within the first line space is selected, but to compensate for this the last address of the cluster will not be selected unless it is completely included within the third space allotted to that cluster.
Single Year Subsample

The selection of the 1957-58 sample was accomplished by taking a subsample of the master sample. It will be recalled that in the master sample every address had a selection probability of \( \frac{3}{1100} \). Since the sampling fraction for the one year sample was \( \frac{1}{1100} \), this requirement was satisfied by taking \( \frac{1}{3} \) of the addresses in the master sample.

In the master sample there were clusters of three addresses. A random number was chosen from among 1, 2, or 3 giving the line of the cluster to be selected. For the year 1957-58, number 3 was the random selection, so the third line of every cluster was included in the sample.

The above discussion applies to all of the areas sampled by City Directories except the western half of Detroit. In the latter case the latest Directory was 1956 with a new Directory scheduled for the fall of 1958. It was decided to list only a one year sample for this area. Because of the greater expense in pulling single lines it was decided to select at the master sample rate (as in the East Side of Detroit), but to list only \( \frac{1}{3} \) of the clusters. The method used was as follows: The pages and clusters were selected as if the complete master sample was to be selected. Then a number equal to or greater than one and equal to or less than three was selected at random. Using this number as a random start and employing an interval of three, every third cluster from the random start was recorded. Thus each line had \( \frac{3}{1100} \) chances of coming into the sample, but here the probability was reduced to \( \frac{1}{1100} \) by taking every third cluster rather than every third line.

Once the sample lines had been selected the actual recording of the addresses proceeded. In general, each line in the cluster was recorded along with a short description of what the line contains, such as a business, household, etc. The decision as to whether the address of such places is a proper
address to give to the interviewer (i.e., whether the address contained DU's) will vary according to a sampler's knowledge of the area. It is usually difficult to ascertain the exact nature of an address from the City Directory; thus where there was any doubt of whether or not an address contained a dwelling unit, an interviewer checked it in the field.

Aside from this aspect, different City Directory addresses were handled in one of two ways. If the listing was an address where only one dwelling unit was expected, the address was listed on a special form and the interviewer was instructed to go to this address and obtain an interview. This type of address has a single line listing or only one name associated with the address, or two names of which the surnames are the same with no indication of different locations in the building. Or it may be a hotel, business or apartment listing with five or more distinct apartments or offices. (Five is an arbitrary dividing point used by the Sampling Section of the Survey Research Center).

The address may require further work if it has two or more different surnames, less than five distinct apartments, addresses with a DU described as "rear", etc. In such cases the address was considered a multiple DU, and the address is listed on a second form.

A different procedure is required in this instance to maintain the equal sampling probability for each address. In these cases, an address occupied several lines in the City Directory; an address of this type was selected when any one of the DU's listed at the address fell into the sample.

For example, the following type of listing frequently occurs:

701 Adam Street
Lazerwitz
Clausen
Sharp

The different surnames indicate that there are three DU's at this address.
Since each of the names occupies a line of print, 701 Adam Street has three chances of coming into the sample thus giving it a selection probability of \( \frac{3}{1100} \). To compensate for this the address is listed on a special form and the interviewer is instructed to interview at every third DU at this address. The probability of a DU being selected into the sample is thereby again reduced to \( \frac{1}{1100} \).

Trailer courts, hotels, and motels are selected into the sample only if the title line is selected. If the title line is selected the interviewer will list all the permanent residents. Subselection is then made from this list.

**The Segment Sample**

The chunk and segment method (a form of area-probability sampling) was used for those areas that were not covered by a City Directory and thus could not be sampled by the City Directory method. The stages of the sampling equation for these areas are similar to the stages in the City Directory equation, and were as follows:

\[
\frac{1}{1100} = \left( \frac{x_1}{2x_1} \right) \left( \frac{y_1}{y_1} \right) \left( \frac{1}{K} \right) \left( \frac{B(2x_1)}{y_1} \right) \left( \frac{1}{1100} \right) \left( \frac{1}{3} \right)
\]

Each term in this equation can be viewed as a stage in the actual sampling process.

The overall sampling rate for the selection of DU's is represented by \( \frac{1}{1100} \). The first stage in the sampling process was represented by \( \frac{x_1}{2x_1} \) and is the probability of selection of PSU's as determined by dividing the number of ODU's in the PSU \( (x_1) \) by the total number of ODU's in the stratum \( (2x_1) \).

The PSU's were then divided into political or other governmental units which have well defined boundaries and fairly reliable population statistics. These may be census enumeration districts, census tracts, precincts, etc. The selection rate was represented by the second term in the sampling equation.
This is the number of ODU's in the subdivision \( (y_i) \) divided by the number in the PSU \( (x_i) \) and represents the probability of a subdivision being selected from that PSU.

These subdivisions were further divided into areas containing approximately thirty DU's which were referred to as chunks. These chunks had roads, streams, and other relatively permanent landmarks for boundaries. These boundaries must be clear and unambiguous in order that the interviewer can be certain of the limits of the area to be covered.

The rate of selection of chunks was represented by the third term in the sampling equation \( \left( \frac{1}{k} \right) \) where \( k \) was such as to yield the desired number of chunks to be selected from the subdivision. The desired number of chunks corresponded to the number of segments expected in the entire stratum and was computed by dividing the total number of ODU's in the stratum by \( \frac{1}{1100} \), and dividing this by four. The factor of four was included in the fraction since we determined that each segment should average about four DU's for the most efficient sampling conditions.

It may be confusing to the reader to be told that chunks were selected according to the number of segments desired. This confusion is dispelled if it is realized that there will be an average of only one segment selected from each chunk when employing an overall sampling fraction of \( \frac{1}{1100} \) in order to keep clustering at a minimum.

In this procedure, as well as in the City Directory method of sample selection, it is convenient to design a master sample which can be utilized for several samples.

The master sample was provided in the Detroit Area Study research by the division of chunks into several segments, only some of which will be used for
each year's sample. Thus each year it is necessary to select a subsample of segments at a rate in accordance with the sample fraction for that year. In 1957-58 the fraction was $\frac{1}{1100}$; the fifth term of the equation therefore gives the sampling rate for the sub-selection of the segments within the master sample for the 1957-58 study.

The fourth term of the equation gave the probability for the sample segments that were expected from a chunk for the master sample of $\frac{3}{1100}$. The actual number of segments were based upon a selection from a continuous list of the expected sample segments in all sample chunks. To do this, a controlled selection procedure was utilized and consisted of a cumulative totalling of the expected number of sample segments required from each subdivision.

The controlled selection procedure was needed because the segment interval will nearly always give values in fractions. For example, $1.648$ may be the number of segments expected from a chunk at a rate of $\frac{1}{1100}$. One procedure of selecting whole segments would be to select them from each of the chunks. Following our example, either one or two segments may be selected, depending upon whether the random start is more or less than $0.648$. Thus for each chunk there is a possibility of a variance of one segment depending upon the random start. If there are fourteen subdivisions it is possible that fourteen more or less segments could have been selected than was required.

By cumulating the total number of segments expected from the chunks and using only one random start, the variation which is a function of the many random starts will be controlled as the fractions are cumulated. The variation of the entire segment sample will not be more than one as a result of the use of this controlled selection procedure, as illustrated below.

As an example, let us suppose that there are chunks $x$, $y$, and $z$ from which a certain number of segments are expected. We would proceed as follows:
For x, two segments are selected since the random start and 1.241 (one plus the random start) are included in the number of segments expected (1.648); for y, ten segments will be used because there are ten values that fall between 1.648 and 11.860: 2.241, 3.241, 4.241, . . . 11.241.

Once the number of segments to be sampled from each subdivision was determined, it was a simple process to: (1) list all the segments in all the sample chunks in a subdivision in a continuous list, giving each segment a measure of size which was a function of the number of dwelling units in the segment divided by four (the average number of DU's we wanted per segment), (2) cumulatively totalling these measures of size, and (3) selecting the proper number of segment measures of size by a systematic sampling procedure. The interval was computed by dividing the total number of segment measures in the chunk sample by the number of segments needed from that subdivision as determined by the controlled selection procedure.

The explanation of the sampling process for the 1957-58 study has aimed at maximum clarity and simplicity of form. For example, we have assumed here that each segment consisted of exactly four DU's. This assumption was not always true. A segment may have had as few as one or two DU's, and as many as thirty. This may have been due to growth in a segment area over a period of two years, or it may have been because a chunk was not divisible into segments of four DU's.
These irregularities were easily controlled by giving a segment a measure of size equivalent to units of four DU's. Therefore, if there were actually eight DU's in the segment, the segment was given two measures of size, etc.

From the cumulative procedure described above, it is apparent that a segment with two measures of size had twice the probability of being selected. Therefore, every DU in that segment had twice the chance of coming into the sample as did DU's in a segment with one measure. Consequently, the interviewer was instructed to interview at every second dwelling unit in those segments with two measures of size, etc.

Summary

We may ask, "How does the above procedure give every adult in our defined sample universe an equal chance of \( \frac{1}{1100} \) of being selected? The best way to understand the answer to this question is to view the sampling process from the perspective of the individual who represents a DU. Let us take a stage at a time and determine the probabilities of John Doe being included in the sample by virtue of his being a member of a PSU, a precinct, a chunk and a segment. Each of these subdivisions has a probability of coming into the sample which is expressed in the terms of the sampling equation. As John Doe lives in each of these subdivisions, he has a probability of coming into the sample which is the product of the probabilities of each of these divisions being selected.

John Doe lives in PSU 1, which has a probability of \( \frac{1}{10} \) of being drawn into the sample; he lives in Precinct 1 which has a probability of \( \frac{1}{5} \) of being selected from PSU 1; he lives in chunk 1 which has a probability of \( \frac{1}{8} \) of being selected from Precinct 1; and lastly, he lives in segment A which has a probability of \( \frac{1}{2.8} \) of being selected into the sample from Chunk 1. Joe, since he represents a DU, will be interviewed since every DU in that segment, assuming it has a measure of size of one, will enter into the sample. The chances of
everyone of the areas in which he lives of being selected are multiplied by each other: \( \left( \frac{1}{10} \right) \left( \frac{1}{5} \right) \left( \frac{1}{8} \right) \left( \frac{1}{2.8} \right) = \frac{1}{1120} \). Therefore we can see that the probability of selection of his household will be \( \frac{1}{1120} \).

**Selection of Respondents in Sample DU's**

Once a household was selected by any of the methods described above, we followed a procedure which selected a given adult in the household as the respondent. This was necessary so that the final sample would not be biased and give the wrong proportions of males and females, housewives and husbands, and so forth.

In addition to the sample address, a Respondent Selection Table was stamped on each interview Cover Sheet. This table was designed to insure that all adults in a selected household had an equal chance of selection. The following two tables are illustrative:

<table>
<thead>
<tr>
<th>Selection Table E₁</th>
<th>Selection Table E₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>If the number of adults in the dwelling is:</td>
<td>Interview the adult numbered:</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

The interviewer first listed all the adults in the selected dwelling unit; this listing was taken in order of age, males preceding females. Each adult was then numbered serially, oldest male first and youngest female last. The interviewer then referred to the selection table and selected the adult whose number corresponded to the number opposite the number of adults in the dwelling unit.
The differential effect of these tables can be seen by comparing $E_1$ and $E_2$, above. If the interviewer determined that there were four adults at a dwelling unit, he interviewed adult No. 3 if using the $E_1$, or adult No. 4 if $E_2$ were on the Cover Sheet.*

The selection of one adult from a dwelling unit regardless of the number of adults residing there introduces a bias which requires theoretical consideration. The bias is due to the differential probabilities of the selection of adults when the sampling unit of selection is a dwelling unit. These probabilities vary with the number of adults within the dwelling unit. In a single adult DU, the probability is $\frac{1}{1}$; in a six adult household, it is $\frac{1}{6}$.

The bias is likely to be insignificant because of two compensating factors: (1) individuals in the same dwelling unit tend to be more alike than different, (2) the number of two adult households comprises 68 per cent of the sample; one and three adult households accounts for an additional 21 per cent and the number of six, seven, eight, and nine adult dwelling units comprises about one per cent. Therefore, the proportion of the sample having extreme probability is relatively small.