A Method of Graphical Analysis

By HELENE C. BATEMAN

Introduction

IN connection with many telephone problems of an economic character, it is necessary to develop methods for making estimates and forecasts of the effects of changes in conditions. When the changes in conditions are such that direct experimentation is impracticable the development of logical methods and bases for estimates involves analyses of past experience in specific situations and, in so far as is feasible, the generalization of such experience. It is the purpose of this paper to describe briefly a graphical method by which complex economic data may be generalized for use in forecasting probable future conditions.

In some problems, it is necessary to determine the effects of changes in a specific situation, the results being applicable particularly to the given situation, and only very generally to other situations. The effect of a change in population upon station growth in a given exchange is an example of such a problem. In other problems, it is practicable to generalize experience so that the results of analyses may be applied, under proper conditions and limitations, to various specific situations. Moreover, it is often necessary to apply a general conclusion to a specific situation because no specific experience is available. An example of this type of analysis is the generalization of results of various rate treatments in different exchanges. In meeting this type of problem graphical methods are utilized to compare experience of a similar nature in various situations. The factors which may be indices of differences in conditions among various situations are studied to determine their relation to the differences encountered. Finally an attempt is made to derive quantitative relationships from the experience analyzed.

The assumption made in utilizing such methods is that the experience in different situations, from which generalizations are to be made, is *essentially similar* in certain respects, and that the variation in the quantitative unit to be estimated is due to varying conditions, as between the different situations, which may be measured in part by quantitative factors. There are, of course, certain types of problems where essential similarity between different situations does not exist or where it is difficult, if not impossible, to isolate quantitative factors sufficiently reliable to form a basis for estimates. On the

other hand, there are many problems to which these methods may properly be applied and in which it is practically impossible to derive a reliable and satisfactory basis for making estimates without some such methods of analysis. Certain economic problems, in particular, because of the impracticability of experimentation and because the complex reactions of a group of individuals are involved, are not adapted to solution by the statistical methods which have proved useful in biometric sciences, but may be dealt with by graphical methods. This has been found particularly true in problems involving local telephone message use, and throughout the following discussion, illustrations are drawn from analyses of this type.

Data

Since the ultimate aim of a graphical analysis of this type is to provide a basis for making estimates, the first step is to determine the estimates which will be required and the type of cases and conditions under which they will be used. In this way the aim and scope of the analysis is clearly defined. The unknown factor (the dependent variable) is to be estimated from certain known factors (independent variables). Various factors, quantitative and qualitative, which might logically appear to be indices of conditions controlling the dependent variable are, therefore, considered.[1] Only factors as to which data are available at the time and place where estimates are to be made are useful as independent variables. It is usually advisable to test a suggested factor by means of any data, even in small amounts, which may be available before a complete body of data is collected. Such preliminary investigations are useful in indicating the scope and detail in which data should be secured. In general the data should:

1. Be adequately representative of the type of cases for which estimates must be made,
2. Be adequate from a sampling standpoint for each situation,
3. Be as nearly homogeneous as practicable, i.e., cases having any outstanding peculiarities should be excluded,[2]
4. Include what appear to be the important factors or indices for each case.

[1] It should be noted that such relationships need not be those of cause and effect. If two factors vary together (as do, for instance, different effects of a common cause) the values of the one which are hard to determine can be estimated from the more easily measured values of the other.

[2] For instance, if estimates are to be made for small exchanges, it would not be advisable to include data from large exchanges in the analysis.

Preliminary Analysis

After the data have been collected and summarized in accordance with the general plan of the study, the graphical phase of the analysis begins with trial setups in which the dependent variable is plotted against each of the independent variables in turn. Such charts are intended only to give a general idea of the types of relationships and to determine which of the factors tested are most closely related to the dependent variable. Factors which do not vary with the dependent variable are not necessarily to be discarded permanently since the effect of one factor may obscure that of another. It is not to be expected that the data plotted on any of these charts will fall along smooth curves. They will probably be widely scattered but in the case of the more important factors a general trend is usually evident.

On the next series of trial charts, several of the more important factors are considered simultaneously. If a qualitative factor is under consideration, separate charts are plotted for the different classes. If these charts are essentially similar, the qualitative factor may be disregarded for the time being and the data considered as a whole. If, however, the qualitative factor appears to influence the relationships in a logical manner the data must be sub-divided and a number of practically independent studies carried on. In fact, the analysis of the effect of a qualitative factor is intended to determine whether or not the data forms an essentially homogeneous whole. If there is a discontinous variable, it is often convenient to hold it constant, i.e., a separate chart may be plotted for each value or group of values of this factor. The factor, which from the preliminary charts, seems most important is usually plotted against the dependent variable. One or two other factors are coded. The codes may be either in colors or symbols or both. The color codes are usually the more easily distinguished and are, therefore, the better for working charts. For final charts, however, color codes are not usually practicable because of the difficulties of reproduction. Both colors and symbols may be used when two coded factors are to be tested simultaneously.

In these preliminary sets of charts, it is well to test as many different factors and combinations of factors as appear logically to vary with the dependent variable. It is usually best, however, to consider not more than three or four independent variables at a time, one plotted against the dependent variable with one or possibly two coded and one held constant on each chart. An attempt to hold constant a greater number will often sub-classify the number of data points so far as to obscure the real trends. Furthermore, the com-

plexity of charts increases rapidly with the inclusion of more variables and makes the analysis and estimating complex and cumbersome.

Fig. 1 is a typical preliminary trial chart from a study of average telephone message use under message rate service. Each data point represents one class of service in a particular exchange. The independent factors taken into account are:

1. Major Service Classifications [3]—held constant since this chart is for one class only.
2. Rank of Service [4]—plotted.
3. Message Allowance—coded.

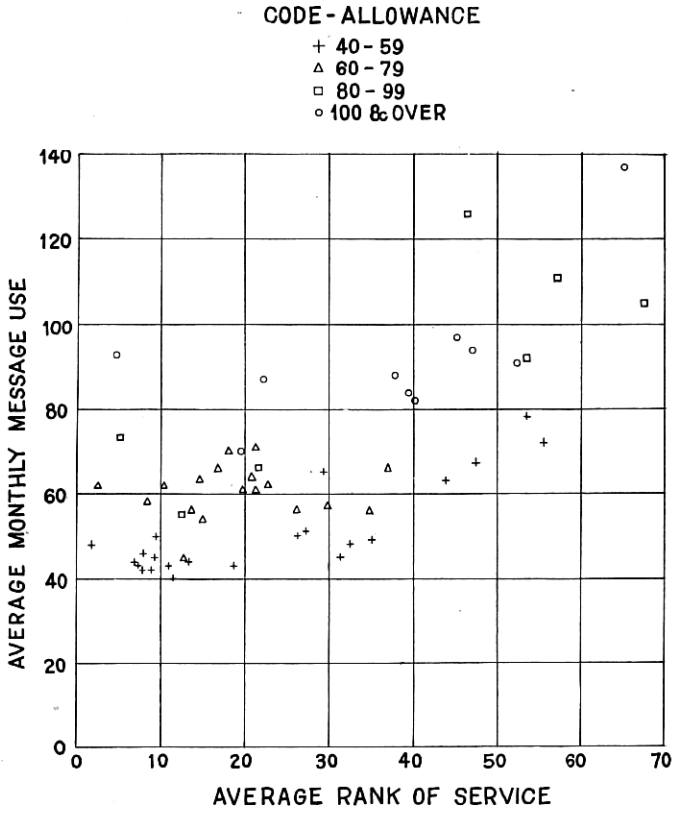

Fig. 1—Preliminary Chart Involving Three Variables

The trend of the relationships between message allowance, rank of service and average message use is fairly well defined on this chart.

[3] Business Main Station, Residence Main Station, P. B. X., etc.

[4] A statistical index indicating the relative ranking of subscribers in accordance with their demands for service.

When several different sets of charts such as are described above have been scrutinized, definite trends will usually be fairly clearly established. It will often be found that while these trends are well defined, nevertheless a number of points may scatter widely. Such points are studied carefully. If, after the original data are checked, the points are found to be correctly plotted, each case is investigated in detail to account for the observed divergence. Sometimes it will be found due to a factor which has not been taken into account, the inclusion of which will often improve the results of the study as a whole. On the other hand, peculiar local conditions or history may give rise to such divergence. These cases are not really a part of the similar group under consideration. If they are sufficient in number and similar with respect to each other they may be studied independently. If not, they are either excluded entirely or given slight weight in the general study. Because of wide differences in problems and material, it is not practicable to describe in detail the process of analyzing such preliminary charts in arriving at decisions as to data and process.

Curve Drawing

The next step is the construction of curves through these data which will truly represent the relationships involved. This can be facilitated by plotting the average values of the dependent variable for all cases having the same values (within certain limits) for all the independent variables.

> On Fig. 2 the data points are the same as those plotted on Fig. 1. The closed symbols which have been added are average points representing the data points of the same symbol. The abscissa of each average point is the mid-point of an interval of rank of service (0–20, 10–30, etc.) and the ordinate is the average of the message uses of all the points falling within that interval.

The average most often used on such charts is the median not only because it is most easily located but because it is usually the most representative, giving little weight to extreme cases. Whatever average is used, it is well to make it a moving average, i.e., covering overlapping intervals such as 20–30, 25–35, 30–40, etc., rather than 20–30, 30–40, 40–50, etc.

These averages serve as a guide for drawing the preliminary curves through the data but the actual data points are considered at the

same time. In constructing the curves, the significance to be attached to any data point depends chiefly on:

1. The number of individual cases on which it is based.
2. The probable degree of accuracy of the data.

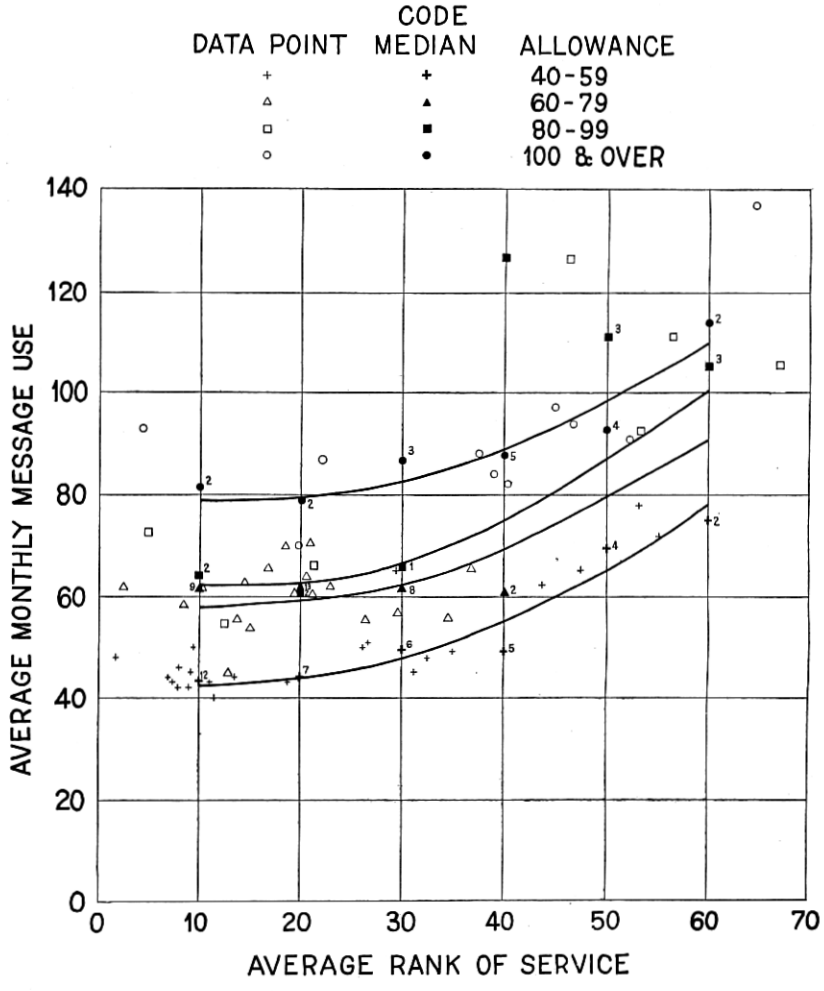

Fig. 2—Preliminary Chart With Averages and Curves

Since these characteristics are considered simultaneously it is usually advisable to depend on judgment using the averages as a general guide, rather than to rely on any formal mathematical system. The first set

of curves is drawn to fit the data as closely as practicable and still be *logical* and *consistent.*

On Fig. 2, the number of data points on which each average is based has been noted as an aid to judgment and a set of rough preliminary curves has been drawn. These, of course, are not necessarily the most accurate curves which could be constructed from these data. A method of progressing to final curves is described below.

Curve Smoothing

The first set of curves constructed from the data may not be an entirely consistent and reasonable family. The relation between different curves on the same chart or between different charts indicates the influence of factors other than the one plotted and must, therefore, be made consistent and logical. The process of transforming the preliminary curves into the final normals is known as smoothing.

The original curves are first studied for reasonableness. Their general shape (whether straight line, convex or concave, having maximum or minimum points, being asymptotic to a certain line, etc.) is, in so far as practicable, determined on logical grounds. If a large majority of the curves, or the curves based on the greatest amount of data, have a certain clearly defined trend, the remainder of the curves are made to conform to this trend, if it is reasonable, at the same time keeping as closely in line with the data as possible.

Each chart will usually have one independent variable plotted against the dependent variable and another independent variable coded. Each curve, therefore, indicates the relationship between the dependent variable and one independent variable for a certain constant value or range of values of a second independent variable. If the relative positions of the curves of a family on one chart are adjusted, the relationship of the coded variable to the other two is altered. The effect of this alteration may be seen by plotting the coded variable against the dependent variable and coding the one which previously was plotted, all values being read from the preliminary curves. This is sometimes called cross-sectioning. The families of curves formed by cross-sectioning are then smoothed until they are reasonable and consistent. In doing this, the original curves are automatically departed from, and when the original curves are replotted from the cross-sections, it may be found that the resulting family of curves is not smooth, consistent or reasonable.

The smoothing process must, therefore, be repeated back and forth a number of times until both sets of curves appear to be smooth, reasonable and consistent families. During this process, it is important that the various families of curves be tested against the original data. If this is not done, it may happen that a series of small changes will accumulate in such a way as to bring portions of the curves outside the limits of the original data. Furthermore, the factor or factors held constant on each chart must not be lost sight of. These factors should be plotted against the dependent variable holding constant

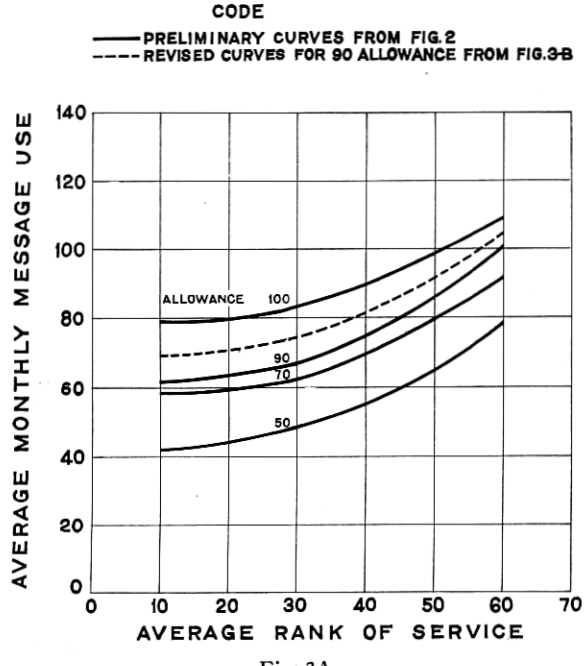

Fig. 3A

all other factors (values being read from the smoothed curves) to see that these relationships also are being made reasonable, consistent and smooth.

The process of smoothing described above is a long and laborious one involving at every step the making of special decisions based upon knowledge of the data and the logic of the situation with regard to the particular problem. Various methods of facilitating the work have, however, been devised some of which are described below. Figs. 3A and 3B illustrate the advantage of having both sets of curves on the same chart with the same scale for the dependent variable so

that when the smoothing process is applied to one family, the effects on the other may be more readily ascertained.

The curves on Fig. 3A are the rough curves which were drawn through the data on Fig. 2. Fig. 3B shows the cross sections of these curves, message allowance being plotted and rank of service coded. It is evident that neither set of curves is a consistent family. Most of the curves of Fig. 3B are irregular instead of being smooth. They might be smoothed considerably either by lowering the points corresponding to a 70 message allowance or by raising those for an al-

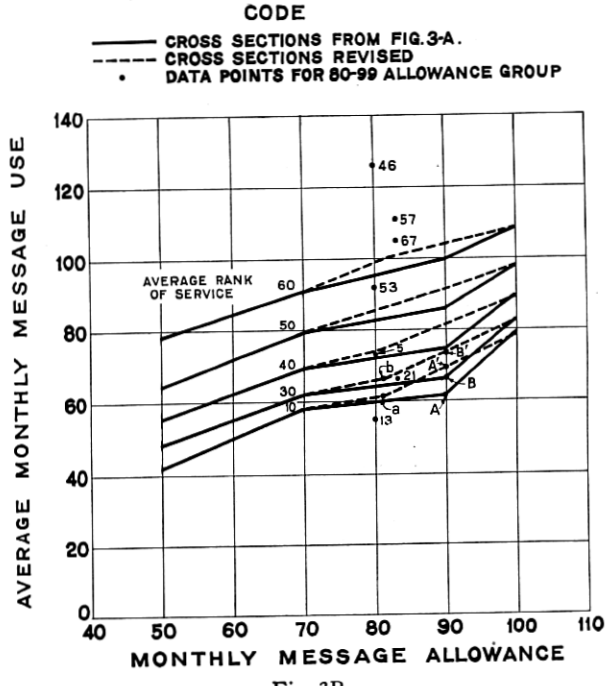

Fig. 3B

lowance of 90 messages. A study of Fig. 3B shows that the curves at 90 allowance are further apart than at any other point. This might be used as an argument either for raising the lower points or lowering the higher. In scrutinizing the data, however, it is found that of the classes of service having from 80 to 99 allowance all have allowances of either 80 or 83. Therefore, the midpoint (90 allowance) is too high to represent the group, or, conversely, the message use plotted is too low for an allowance of 90 messages. In order to have a guide in the amount of shifting necessary, data points for the actual message allowances of the 80–99 group have been plotted

on Fig. 3B, and the values formerly entered at 90 (points A, B, etc.) entered at 81½ (points a, b, etc.). With these points and those at 100 as a guide, new values for 90 allowance have been estimated (points A′, B′, etc.). The shifting of a point up or down on Fig. 3B results in shifting the corresponding points of the other family (Fig. 3A) the same distance in the same direction resulting in the dotted curve. There is much more smoothing necessary to make Figs. 3A and 3B satisfactory and reasonable, but by proceeding in the manner just described, taking into account the appearance of the curves, the logic of the situation and the original data, a smooth and consistent family of curves can finally be evolved.

Another excellent method of smoothing involves the use of a three dimensional figure. Just as a plane surface gives a complete representation of two variables and a partial representation (by coding) of a third, so a three dimensional system can be used to give a complete representation of three variables, and a partial representation of a fourth. It also aids greatly in smoothing simultaneously. A device for three dimensional representations consists of a plane surface marked off with rectilinear coordinates and having at frequent intervals holes into which pegs can be set. The pegs also have coordinate markings. The values of two variables, then determine the point at which the peg is set and the value of the third determines a distance along the peg. The point is marked by a small rubber ring which fits around the peg. The values of a fourth variable may be coded by using rings of different colors. When the device is used for smoothing curves involving only three variables, the data points may be indicated in one color and the smoothed values in another. The data points, remaining constant as the smoothed curves are shifted, form a continuous check on the divergence of the smoothed curves from the data. This is an automatic process of cross-sectioning. When the position of a point is changed, the effects of the change on the various relationships are seen by studying different aspects of the setup. This device gives the best results with discontinuous variables (such as message allowances, rates, etc.) as the pegs can then be set in at regular intervals without resorting and regrouping the data. It is also especially valuable when one of the variables is a complex factor (such as distribution of development among more than two classes of service) which cannot easily be represented by one curve.

Fig. 3C illustrates such a setup with the revised curves of Figs. 3A and 3B. The independent variables, message allowance and rank of service are represented by the rectilinear co-

ordinates of the plane surface. Message use is indicated by the rings on the upright pegs.

In general, the various steps in the analysis leading up to the final[5] or normal[6] relationships require continuous exercise of judgment. The problem is never one of securing simply curves of "best fit" to the data. It is broader, more fundamental and much more involved

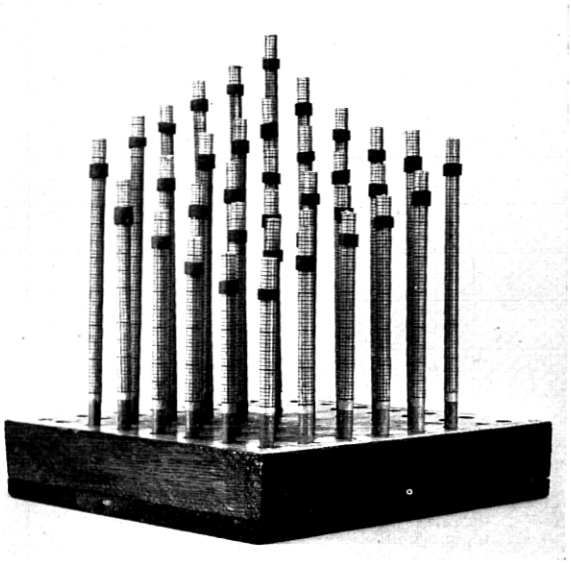

Fig. 3C

than this. It requires a combination of logic with the data that results in normal relationships which fit the data and at the same time are reasonable. It is necessary to consider such questions as the following: Why do the data indicate this relationship? As a generalization, is such a relationship reasonable? What should be the character of this relationship? Should it be a straight line, concave up or concave down? Particular attention is given to the reasonableness of maxima or minima points and to points of inflexion when indicated by the data. It is only by considering such fundamental questions that a sound basis can be established for building up normal rela-

[5] Final in a relative sense. In economic studies of this type involving human reactions and relationships normal relationships are never final in an absolute sense.

[6] The term "normal curve" is used throughout this paper to designate a final curve from which estimates are to be made. A normal distribution curve in this sense may or may not be "normal" in the statistical sense of an evenly balanced bell-shaped curve.

tionships which will be a true generalization of experience. The importance of dealing with economic problems in this way can hardly be over emphasized. It is a recognition of the complexities which are inherent in problems involving human reactions and the dangers of untrue generalization if rigorous and more or less inflexible methods of analysis are utilized.

Final Results

The result of the smoothing process is the development of a consistent series of charts and curves by means of which the value of the dependent variable may be estimated from the values of the various independent variables.

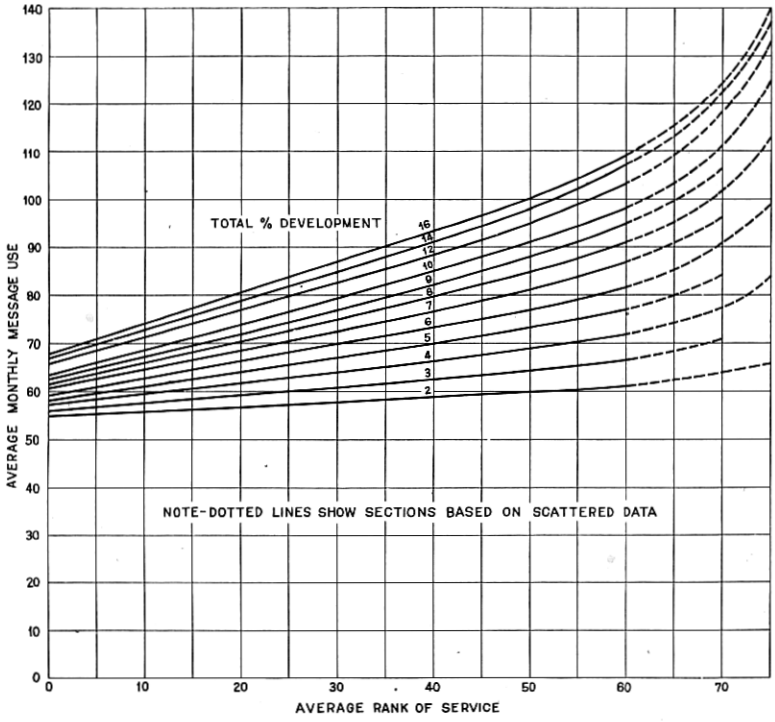

Fig. 4—Final Family of Curves

Fig. 4 is an example of such a chart for estimating average message use. Charts of this type may be used under proper conditions for estimating either an actual value which is unknown (such as average message use under an existing rate schedule) or the value which may be expected to result from some change (such as average message use under a proposed rate schedule).

After deriving a series of final charts estimates are made of the value of the dependent variable for all the cases on which the study was based. Consideration of the differences between the estimated and the actual values is an excellent general criterion of the accuracy of the normals. In general, the positive deviations should be approximately equal to the negative both in number and in the sum of their numerical values. If either positive or negative deviations are decidedly predominant, it is probable that the general level of the normal curves is too low or too high.

When the deviations (without regard to sign) are plotted as a frequency curve, the curve should be fairly smooth. It need not be and usually is not a bell shaped curve, but if there are sudden and decided breaks, it is probable that either certain portions of the data have not been given proper consideration or that the data were not originally essentially homogeneous. The cumulative frequency curve based on the deviations makes possible the easy reading of the median or probable error of estimate. The probable error may be used as a general criterion of the value of future estimates made from these normals and the ratio of the probable error to the median value of the dependent variable forms a basis for comparison of the relative accuracy of different sets of normals.

The deviations (sign being taken into account) when plotted against the various factors included in the study should be fairly evenly scattered and show no trend or relationship. If a consistently occurring variation is discovered between the deviations and any of the independent variables it indicates that the relationship of that variable to the dependent variable has not been properly taken into account in deriving the normals. If this variation appears in connection with the dependent variable, it indcates that some of the curves are not of proper shape. For instance, if a straight line is fitted to data having a decided non-linear trend, the errors plotted against the dependent variable will fall along a well defined U-shaped (or inverted U-shaped) curve.

Additional information may also be obtained by plotting the deviations against factors not included in the study. Relationships will sometimes become apparent which previously were obscured by the effect of the more important factors. The influence of such factors may account for seemingly abnormal cases and their inclusion would tend to reduce the mean and to a lesser extent the median deviation.

Frequency Distributions

Normal curves, such as those described above, form a basis for estimates of an *average value* for a group of items comprising a unit such

as has been utilized in developing the study. In many instances, however, it is necessary to know not only the average value but also the distribution of items about that average.

> Thus, the normal curves of the type of those shown in Fig. 4 serve as a basis for estimating the average message use of all subscribers to a given class of measured service in a given city. Additional curves are, however, required for estimating the distribution of subscribers by message use.

The basic principles governing the derivation of normal curves are the same whether these normals be concerned with averages or with distributions. The detailed methods involved are, however, quite different because of the inherent differences in the material. An average can be expressed in one arithmetic term which can be plotted against other factors. A distribution, on the other hand, is a complex entity which may itself be expressed as a curve but which obviously cannot be measured by an index to be plotted against other variables without losing sight of certain detailed characteristics of the distributions. The procedure and methods described above for deriving normal curves are modified somewhat in the derivation of normal distribution curves. Some of the methods which have been found advantageous for these analyses are described below.

The first step in the analysis is usually to plot the actual detail and cumulative distributions for each group of items and to compare the various distributions in order to determine points of similarity or difference. For purposes of comparison, percentage distributions are used, i.e., the per cent. of total items rather than the actual number occurring in each interval is plotted. With homogeneous material it will usually be found that when plotted to the same scales the detail frequency curves are all of the same general shape but differ in three primary characteristics.

1. The spread or extent of variation.
2. The location of the peak or point of maximum frequency.
3. The concentration of items in the peak interval.

These characteristics are, however, interrelated and to a certain extent related to the average.[7] Other things being equal, it might be expected that:

1. The greater the average the greater the spread.
2. The greater the spread the less the concentration at the peak.
3. The higher the peak the nearer it will fall to the average.

[7] Throughout this section the term "average" is used in the sense of arithmetic mean.

Since the average is of much importance in determining frequency curves, it will usually be found that differences will be reduced if the curves are plotted with each interval of the horizontal scale as a per cent. of the average instead of an actual value. For example:

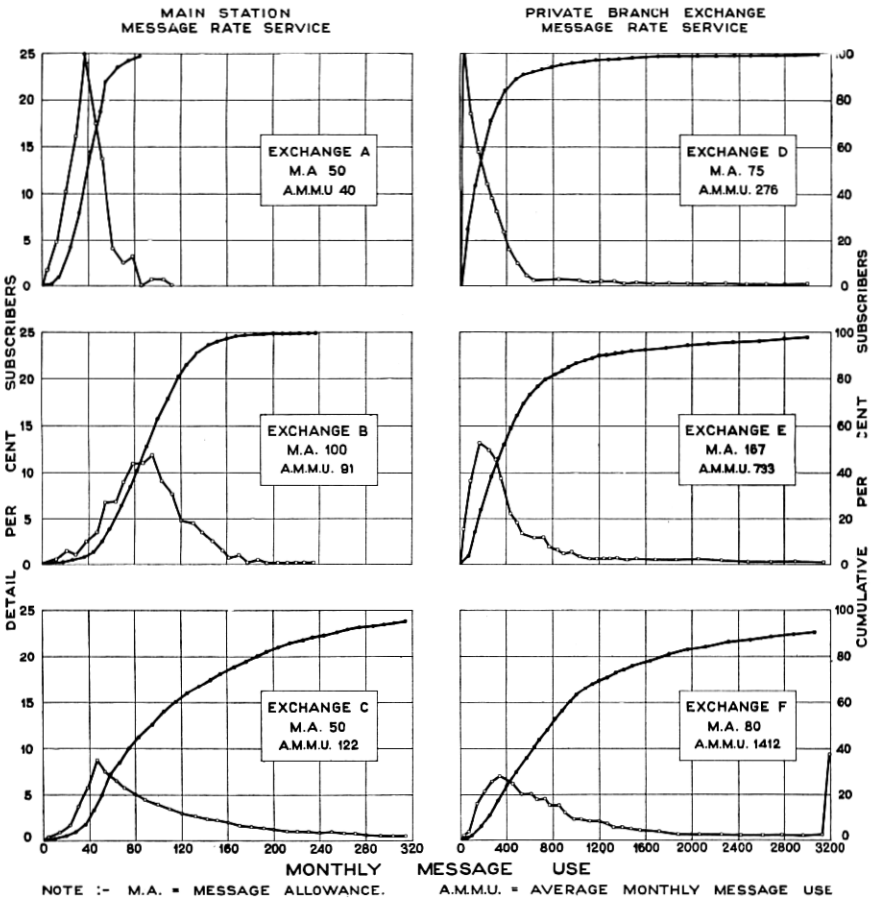

Fig. 5—Sample Distributions of Subscribers by Absolute Message Uses

Fig. 5 illustrates distributions of subscribers by message use plotted in terms of actual values for different classes of message rate service. On each chart the average message use is indicated. It will be noted that, in general, the greater the average message use, the greater the spread of the curves, the less the concentration at the peak interval and the less marked the correspondence of the peak with the average value. On Fig. 6 the frequency curves

have been replotted using instead of each actual message use interval the per cent. that the message use is to average message use. The result of this statistical process is to make the spread of the curves and the height of the peaks much more nearly uniform.

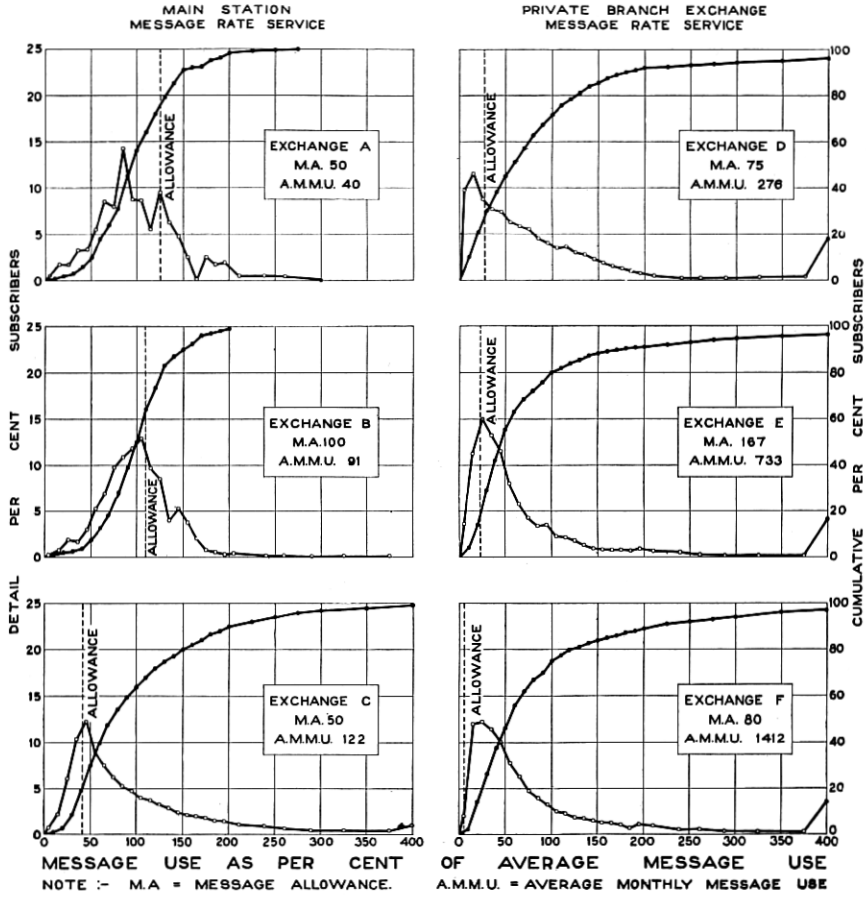

Fig. 6—Sample Distributions of Subscribers by Message Uses in Per Cent. of Average

In some instances, the frequency curves plotted with each interval expressed as per cent. of the average may be so similar for the different groups that satisfactory normals may be derived from this setup alone without including any other factor. This appears to be the case for the distribution of P. B. X. subscribers by message use as illustrated on Fig. 6. It is necessary, however, to test whether or not the full effect of the average on the distribution has been eliminated. This

may be done on a detail basis by plotting a series of charts showing the relation between the absolute amount of the average and the per cent. of cases falling within a given message use interval (expressed as per cent. of the average). On a cumulative basis the per cent. of cases falling below a given per cent. of the average is plotted against the average. For example:

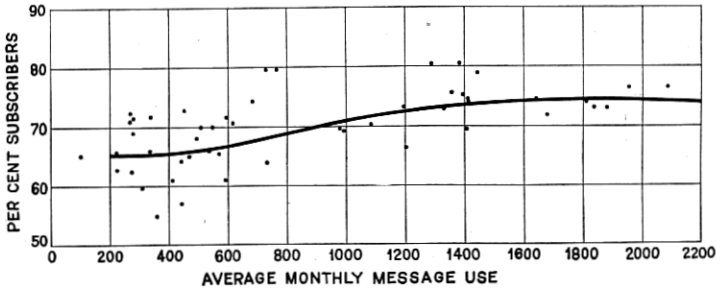

Fig. 7—Preliminary Chart of Cumulative Series

Fig. 7 shows the relationship between average message use and the per cent. of subscribers using less than 100 per cent. of the average message use for a given service classification. Each plotted point represents the reading from the cumulative curve for a different exchange. It is evident that the two factors vary together.

Curves similar in type to that shown in Fig. 7 are constructed on each of the charts of the detail and cumulative series. The curves of each series are smoothed by cross-sectioning and developed into consistent and reasonable families.

In connection with the smoothing of the cumulative series a method described below has been found useful. This method can be used with any setup of three variables but is simplest in setups of a cumulative type which have no maxima or minima within the limits of the curves. To simplify the explanation of the method the cumulative distribution of certain subscribers by message use is referred to as an example.

Fig. 8 shows preliminary curves representing the relationship between average message use and the per cent. of subscribers using less than the various per cents. of the average message use from 10 per cent. to 500 per cent. of the average. These curves are derived from a series of charts similar to Fig. 7 for different message uses. Cross sections of the family of curves of Fig. 8 give a series of cumu-

lative curves. The successive curves of this cumulative series have been plotted in Fig. 9 at regular intervals apart, the intervals being the same distance as the average monthly message use scale of Fig. 8. The horizontal scale of Fig. 9 used in plotting these cumulative distributions must therefore be movable so as to apply in turn to each

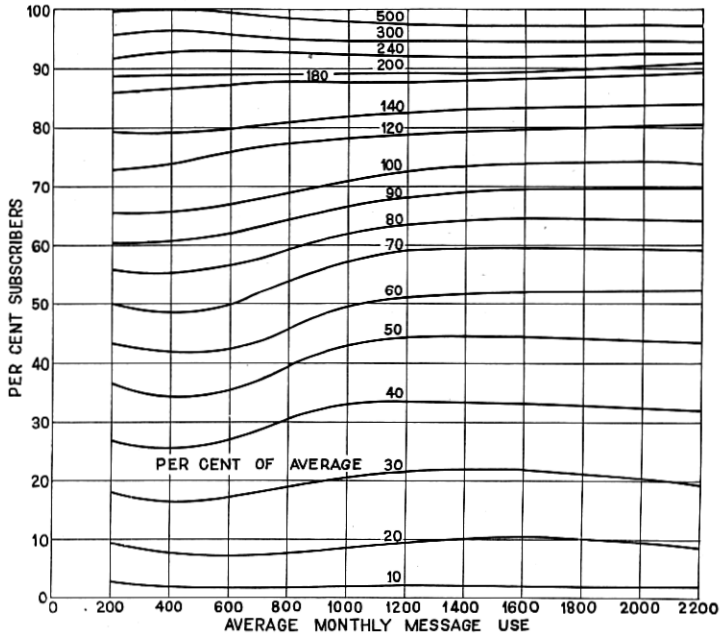

Fig. 8—Preliminary Family of Cumulative Curves

of the cumulative curves. With the cumulative curves plotted, the family of curves on Fig. 8 have been drawn in on Fig. 9, the curves representing the various message uses being exactly the same as those of Fig. 8 except that the method of drawing the cumulative curves has automatically shifted successive curves of Fig. 8 further and further to the right. It follows from the methods which have been used in constructing Fig. 9 that any given cumulative curve must intersect each curve of the other family somewhere on the vertical line corresponding to the message use (expressed as per cent. of the average message use) represented by that curve. For instance, the cumulative curve for 1000 average message use (indicated by A) must intersect the curve representing 30 per cent. of the average message use on the vertical line corresponding to a co-

ordinate of 30 on the horizontal movable scale when the zero point of the horizontal movable scale falls at 1000 average message use on the fixed horizontal scale. The point of intersection described in this illustration is indicated on Fig. 9 by P. This characteristic (inter-

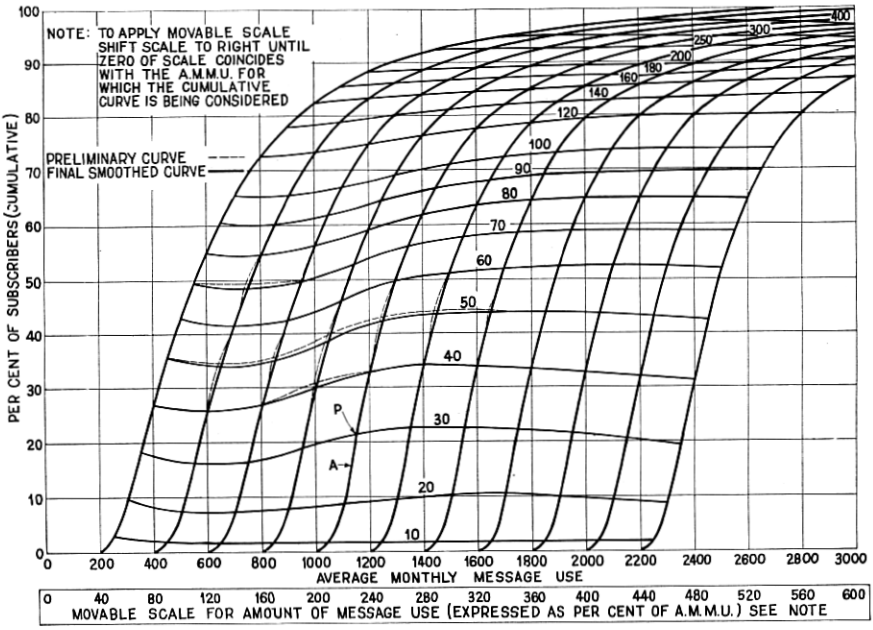

Fig. 9—Simultaneous Smoothing

sections falling on given vertical lines) forms the basis for smoothing the two families of curves simultaneously. A point of intersection may be shifted vertically but cannot be shifted horizontally since it must fall somewhere on a definite vertical line. Dashed lines (- - - -) on Fig. 9 indicate the manner in which a few of the points have been shifted in smoothing.

A family of cumulative curves may appear easier to smooth than the corresponding family of detail curves. On the other hand, the detail curves give, in some respects, a more vivid picture of the outstanding characteristics of the distributions than do the cumulative, and certain important characteristics of the distributions may be more easily studied on a detail basis.

It is important, therefore, that both series be taken into account in deriving final normal distribution curves. For the detail series, charts are plotted showing the relationship between the amount of

the average and the per cent. of cases falling in a particular interval (expressed as per cent. of the average). Fig. 10 is such a chart for the interval 80–90 per cent. of the average message use.

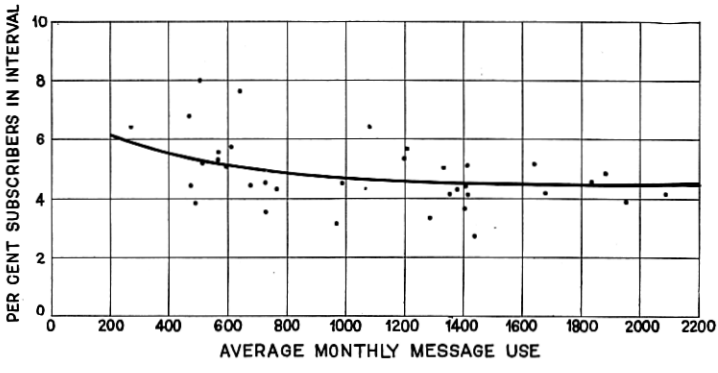

Fig. 10—Preliminary Chart of Detail Series

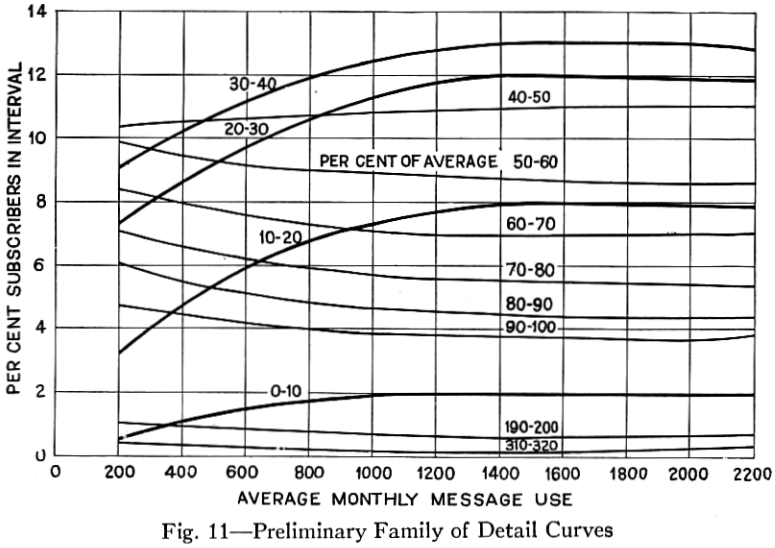

Fig. 11—Preliminary Family of Detail Curves

Cross-sections of a family of curves such as those on Fig. 11 give a series of detail frequency curves. Further smoothing may be facilitated by a study of these curves. As an aid in this process of smoothing it is desirable to determine the normal location of the peaks and the normal proportion of cases occurring in the peak interval, as these are important characteristics of such curves. These normal values may be determined by plotting these factors against

the average as shown on Figs. 12 and 13. For this data, it is noted that, on an absolute message use interval basis, the greater the average the greater the abscissa of the peak value. However, with an increase in the average, the abscissa of the peak interval on an absolute basis

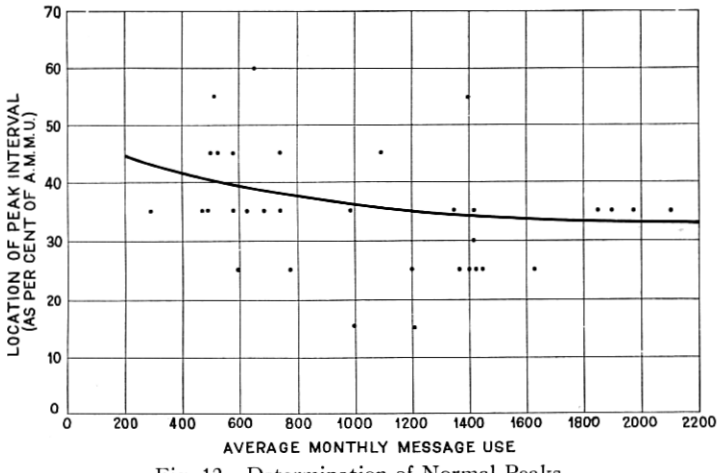

Fig. 12—Determination of Normal Peaks

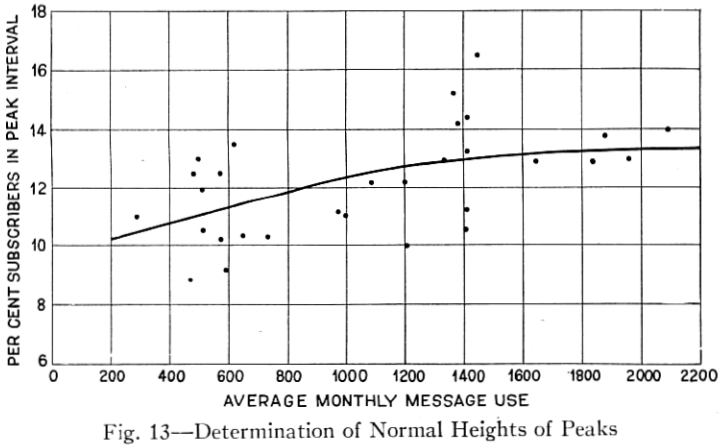

Fig. 13—Determination of Normal Heights of Peaks

does not increase as rapidly as does the average, because large users increase their usage relatively more than small users. Therefore, when the intervals are plotted in terms of per cent. of the average, it will be found that the greater the average the less the abscissa of the peak. For the same reasons, the peak interval expressed as per cent. of the average becomes relatively wider as the average increases, and the height of the peak increases as illustrated on Fig. 13. With

the location and height of the peaks normally determined, the process of constructing preliminary curves for the various intervals of the detail series is, in many cases, considerably simplified. These preliminary curves are then cross-sectioned and smoothed into a consistent family.

Finally the smooth curves of the cumulative and detail series are checked with each other and averages are computed from these curves as a check against the assumed average. When the minor discrepancies disclosed by these checks have been corrected, normal curves are plotted and comparisons are made with the actual distributions. Further adjustments may then be necessary.

Additional Factors in Deriving Normal Distribution Curves

The smoothing processes described above give a series of normal [8] distribution curves taking into account completely the effect of the amount of the average upon the distribution. In some cases, however, it will be found that some outside factor has also a decided effect upon the distribution.

When the effect of an outside factor is apparent it may be necessary to derive a series of normal distribution curves, each curve corresponding to a constant value of the factor under consideration. If this is done, the curves are smoothed by cross-sectioning and the various other methods described above so as to form a consistent and reasonable family. The type of the final family derived will, however, depend largely upon the character of the relationships developed during the smoothing process. For instance, in the case of main station message rate service, a series of distribution curves was plotted, one for each message allowance. In the course of smoothing these curves it seemed reasonable that there might be a relationship between the type of distribution and the proportional relationship of average message use to message allowance. That is, with an annual message allowance of 600 and an average annual message use of 400 the distribution of subscribers by amount of message use might be similar, on a proportional basis, to the distribution of subscribers under an annual message allowance of 900 with an average annual message use of 600; or under an annual message allowance of 1,200 with an average annual message use of 800. This idea was tested by use of the various sets of normals which had been derived for the different message allowances and was found to hold so closely that this pro-

[8] See Note 6.

portion factor (ratio of the average message use to the message allowance) might be used in deriving a revised setup of normal distribution curves.

In certain cases it may be found that some expedient such as that described above may be used to eliminate or take account of the effect of an outside factor. Whether this is done or a separate set of curves is derived for different values of that factor, the process of deriving the detail and cumulative distribution curves would in general be the same as that described above.

Some of the processes involved in studying averages and distributions of subscribers by message use have been described because they are typical and illustrate what have been found to be satisfactory methods of analysis for problems of this type. It is clearly impossible, however, to set up any rigid methods for such studies. Any economic problem which permits of analysis by these methods must be treated in the manner best suited to the data available, the purposes of the study, the degrees of accuracy necessary, etc. Where these methods can logically be employed the results obtained, an important part of which are the background and sidelights, on the problem, disclosed during the process of building up the normal relationships, will generally be found superior to those obtained through the use of more rigorous methods.

Application

Before final results are obtained, there will naturally be developed by those concerned in the study a very definite conception of the field of their usefulness and their limitations. It is important that a knowledge of these limitations be extended to those who may have occasion to use the results. Given a set of smooth curves from which quantitative estimates can be made, there is a great temptation to make estimates under any and all circumstances, and often to give such estimates an undue appearance of accuracy. The final results are merely the general expression of the information contained in the original data logically developed according to the knowledge and judgment of the investigator. It is always necessary in applying such results to consider the effect of special and local conditions. Where it is known that actual conditions in a specific case are far from normal, it is often possible to estimate the effect of a proposed change by applying differences based on the normal experience.

Care must also be taken in extrapolation estimates, i.e., estimates where the value of one or more factors lies beyond the limits of the original data. Such estimates, of course, are always subject to con-

siderable error. In other cases it may happen that some part of the data necessary for making a complete estimate is not available. It may be practicable, however, to approximate the required information and make a rough estimate which may be more accurate than the alternative of basing the estimate upon less complete analysis.

In applying the results of such analyses, satisfactory conclusions can be reached only if due consideration is given to the following points:

1. The quantitative readings from the normal curves.
2. All the qualitative relationships developed in the course of the analysis.
3. Any additional data available for the particular case or cases in question.
4. Any peculiar special conditions known to exist for that case or which probably exist because of comparison with similar cases.
5. Changes affecting general levels since the date of the study.

It follows that the making of such estimates cannot be left in inexperienced hands any more than can the progress of the original study. Good judgment and a complete knowledge of the problem are of paramount importance both in making the general analysis and in the application of results to specific problems.

To those accustomed to working in the more exact fields of physics, chemistry, etc., it will undoubtedly appear that the methods described above may be inexact and unsatisfactory. Undoubtedly, the average errors of estimate are considerably greater than would be allowed in fields where more exact data are obtainable. Yet the reason for this lies rather in the material itself than in the methods of dealing with it. An economic quantity is extremely complex and difficult to estimate because it is usually dependent upon the action of hundreds or thousands of individuals each one of whom is influenced by individual needs and desires which at best can only be partially measured by such quantitative factors as reflect variations in these needs or desires. Estimates of such quantities are necessarily subject to a relatively high degree of error if comparisons are made with the fields of physical science. Yet such estimates must be made and the problem is to make them as accurately as practicable. Judged from this standpoint, experience indicates that such analyses are an important aid in connection with certain phases of many economic problems.