COWLES FOUNDATION DISCUSSION PAPER NO. 47

Note: Cowles Foundation Discussion Papers are preliminary materials dirculated privately to stimulate private discussion and critical comment. References in publications to Discussion Papers (other than mere acknowledgment by a writer that he has access to such unpublished material) should be cleared with the author to protect the tentative character of these papers.

An Analysis of the Distribution of Wages and Salaries in Great Britain

T.P. Hill

January 15, 1958

An Analysis of the Distribution of Wages and Salaries in Great Britain

By

T.P. Hill

In this paper methods developed for the analysis of experimental data are applied to the analysis of income data derived from a cross-section survey. Problems of analysis which are generally avoided by the use of efficient experimental designs are seen to arise in acute forms when the data are obtained from cross-section surveys.

In the second part of the paper, distributions of expected incomes based on a simple model of income generation are presented and a short analysis of the residual incomes is carried out. This discussion throws light on the role of some important demographic variables in shaping the distribution of employment incomes and on the validity of certain hypotheses which have been advanced to explain the distribution of income.

I. A Description of the Data

The analysis in this paper is based on a distribution of annual wages and salaries for men in Great Britain. The distribution was obtained from the national survey of incomes and savings carried out by the Oxford University Institute of Statistics in Spring 1954 and relates to the financial year April 1953 through March 1954. Although the basic sampling unit for the survey did not consist of an individual person, information was collected about the income from employment of all men interviewed which enabled the distributions and tabulations required for the following analysis to be made. A description of the coverage, methods and reliability of the Oxford Savings Surveys can be found in a series of articles in the Institute's Bulletin from 1953 onwards.*

---

* See in particular [2].

Individuals who had not been working full-time, or who had not been in paid employment for at least 46 weeks out of the year, were excluded from the analysis; the number of men remaining in the sample was 1,414 but not all observations carried equal weight because of the sampling methods used. By roughly standardising the time period to which the wages and salaries refer, they can be regarded as equivalent to annual rates of remuneration for the individuals concerned. The distribution obtained presumably exhibits less variability than a corresponding distribution referring to a single month or week, since most of the short-run fluctuations in wages or salaries will tend to be averaged out when the longer time interval is chosen. This is particularly important for manual workers whose wages

may vary considerably from week to week as a result of overtime or piecework, for example. The actual distribution to be examined in this paper is shown in Table I.

(See Table I)

Although the distribution is based on a sample, it displays no obvious irregularities which might be attributed to sampling fluctuations; the distribution is unimodal with the usual marked positive skewness. The few very low incomes are largely to be explained by the retention in the sample of individuals who were not in receipt of full rates of pay because they were still undergoing some form of training -- apprentices or articled clerks, for example. While students undergoing full-time education were not, of course, included in the sample, all full-time paid male employees aged 18 years or over were retained even though they might still be in the course of training. Compared with most income distributions, that in Table I displays a very low degree of inequality; the conventional/Lorenz coefficient is only 21%, while the Pareto coefficient is as large as 3.3. The Pareto distribution seems to give a very good fit to the upper tail of the distribution, although it is impossible to determine from so small a sample whether it would continue to do so in the very highest reaches of the distribution of wages and salaries. Since a deliberate attempt has been made to restrict the analysis to a single type of income and a relatively homogeneous group of income recipients, it is not surprising, of course, that the degree of inequality is much less than for more general distributions of income.

II. The Method of Analysis

The objective of the method of analysis is to exhibit systematically the effects of a variety of factors such as occupation, age, or industry on the level of an individual's remuneration. Since the factors in terms of which the variation in income

is to be explained are essentially attributes rather variables, conventional regression techniques are not appropriate and resort has to be made to techniques of analysis developed primarily in connection with the statistical analysis of experimental data where identifiable but non-measurable factors are liable to influence the outcome of particular experiments. Since these techniques have seldom been applied extensively to economic data, the appropriate methods of analysis will be briefly described first. Essentially, this involves specifying the underlying model or structural relationships and the corresponding estimation procedures.

The first difficulties to be faced are problems raised by the classification of the data according to various criteria. For example, there are numerous ways of classifying individuals by their occupation and the number of separate occupations distinguished is obviously arbitrary and subjective. The investigator has to be guided by pragmatic considerations and to decide on a limited number of occupational groups, in our case nine. However, the logic underlying the subsequent analysis is really based on the assumption that an individual can be allocated to one out of a set of mutually exclusive (occupational) categories on the basis of some specific and unambigous characteristic. If is tacitly assumed that all members of a particular occupation are literally homogeneous in respect of their possession of some specific attribute, and that this attribute exerts a distinct influence on the level of earnings of all individuals with that attribute. This assumption can evidently only be justified as an approximation to reality which acquires greater validity as the number of categories separately distinguished is increased. Similar remarks apply to the other classifications used in this study. Although age is a continuous variate, individuals have been classified into six age groups which are treated similarily to groups within other types of classification. One reason for adopting this procedure

is that it avoids specifying the particular form of the functional relationship between income and age, which appears to be rather complex; it also enables age to be handled in the same way as other factors.

The method of analysis to be adopted can be illustrated by examining a situation in which only two principal factors, such as occupation and age, are assumed to influence an individual's earnings. A simple model is set up as illustrated in equation (1).

$$y_{ijk} = m + a_i + b_j + e_{ijk} \qquad i = 1,2,\ldots r,\ j = 1,2 \ldots s\ . \qquad (1)$$

where y_{ijk} denotes the income of the kth individual in age group i and occupational group j ;

m denotes a parameter common to all individuals;

a_i denotes a parameter common to all individuals in age group i ;

b_j denotes a parameter common to all individuals in occupational group j ;

e_{ijk} denotes a random variable with zero mean and finite variance.

Any set of observations can be arranged in the form of a two-way classificational table with rs cells. When the observations are selected at random from a given parent population the numbers of observations falling within particular cells will be proportional to the proportions in which the corresponding observations occur in the population, apart from sampling fluctuations. Let us denote the probability of obtaining an individual of age i and occupation j by the symbol p_{ij}, where $\sum_i \sum_j p_{ij} = 1$. The expected numbers of observations within particular cells of the sample are therefore given by equation (2):

$$E(n_{ij}) = Np_{ij} \qquad (2)$$

where N is the total number of individuals in the sample. The expected level of income for all individuals of age i will be denoted by the symbol $E(y_{i.})$, where the use of the period in place of the subscript j indicates that the expectation embraces individuals in all s occupational groups within age group i. On the basis of equations (1) and (2) we may write:

$$E(y_{i.}) = m + a_i + \frac{\sum_j^s (p_{ij}b_j)}{\sum_j^s (p_{ij})} \qquad (3)$$

It should be noticed that the absolute level of the set of parameters a_i or b_j is indeterminate since either set of parameters can be increased or decreased by a constant amount provided that the overall parameter m is adjusted by an equal amount but in the opposite direction. The difference, however, between any pair of age (or occupation) parameters is uniquely determined by the specification of the model. The corresponding difference between the expected incomes for two age groups, say 1 and 2, is given in equation (4).

$$E(y_{1.} - y_{2.}) = a_1 - a_2 + \left\{ \frac{\sum_j^s p_{1j}b_j}{\sum_j^s p_{1j}} - \frac{\sum_j^s p_{2j}b_j}{\sum_j^s p_{2j}} \right\} \qquad (4)$$

It is clear, therefore, that the difference between the average incomes for any pair of age groups is generally a biassed estimate of the difference between their age parameters, and similarly for occupation groups. This bias is independent of the

size of the sample and cannot be reduced by increasing the sample size. However, if the relative occupational distribution within each age group happens to be the same for all age groups, the bias is eliminated. More specifically, if $p_{ij} = p_i p_j$ for all i and j, the expression in braces in equation (4) vanishes identically for every pair of age groups, and also in the corresponding expression for occupation groups.

If the object of the investigation were principally to examine the effects of age and occupation on income, it would be better to select the sample along quite different lines. By analogy with efficient experimental designs, it would simplify the analysis considerably if equal numbers of observations were selected (at random) within each age-occupation sub-group. If the population could be suitably stratified prior to sampling, this would be a simple operation. However, the requisite sampling frames are usually non-existent in practice, and in any case actual cross-section surveys are taken with many different objectives in view, which generally leads to simple random sampling with certain modifications which are not important here. In general, therefore, it can be concluded that whenever the data arise from cross-section surveys there will infallibly be unequal numbers of observations, and sometimes very unequal numbers, in the cells of a classification table of the kind considered here. Given the validity of the underlying model, the appropriate method of estimating the parameters will now be reviewed briefly.

Best linear unbiassed estimates of the parameters in equation (1) are obtained by the use of conventional least-squares estimation procedures; if it also happens that the e_{ijk} s are normally distributed, least-squares estimates yield the maximum likelihood estimates. The normal equations for estimating the parameters are given as follows:

$$\sum_i \sum_j \sum_k y_{ijk} = N\hat{m} + \sum_i n_{i.}\hat{a}_i + \sum_j n_{.j}\hat{b}_j$$

$$\sum_j \sum_k y_{ijk} = n_{i.}\hat{m} + n_{i.}\hat{a}_i + \sum_j n_{ij}\hat{b}_j \qquad (i = 1,2, \ldots r) \qquad (5)$$

$$\sum_i \sum_k y_{ijk} = n_{.j}\hat{m} + \sum_i n_{ij}\hat{a}_i + n_{.j}\hat{b}_j \qquad (j = 1,2, \ldots s)$$

where n_{ij} denotes the number of individuals in age group i and occupation group j as previously,

and $n_{i.} = \sum_j n_{ij}$ and $n_{.j} = \sum_i n_{ij}$.

These equations are not independent, which reflects the indeterminacy in the levels of the sets of parameters a_i and b_j commented on above. In particular, the first equation in (5) is equal to a half the sum of the remaining $(r + s)$ equations. The neatest way of resolving the indeterminacy is to add the following specifications:

$$\sum_i n_{i.} a_i = \sum_j n_{.j} b_j = 0 \ , \quad \text{which implies} \quad m = \bar{y} \qquad (6)$$

These seem desirable properties for the sets of parameters to possess; the differences between parameters in any set are, however, uniquely determined by the normal equations and are not affected by these additional specifications.

One special case needs to be noted. This arises when the proportionate distribution of the observations within the rows (and columns) of the corresponding two-way classification table is constant. Formally, this implies the following relationship among the n_{ij} s :

$$\lambda_1 n_{1j} = \lambda_2 n_{2j} = \lambda_3 n_{3j} = \ldots = \lambda_r n_{rj} \quad \text{for every} \quad j \qquad (7)$$

where the λ_is comprise a set of constants.

Given a distribution of observations in the sample which conforms to (7), the terms involving $\hat{b}$ can be eliminated from the second set of equations (r in total) in (5), while the terms involving $\hat{a}$ can be eliminated from the third set of equations (s in total) in (5). In these circumstances, the estimates of the a's are independent of the estimates of the b's, and vice versa, and the distribution of the observations is said to be orthogonal. Care is usually taken to ensure orthogonality in the design of experiments, but non-orthogonality will almost invariably arise when the data are taken from cross-section surveys. Of course, if $p_{ij} = p_i p_j$ in the parent population, the distribution of observations in a random sample from that population will be at least approximately orthogonal. Finally, if all the λ_is in (7) happen to be equal to unity, and there are equal numbers of observations in all possible sub-cells, the analysis is especially easy. In general, whenever the distribution of the observations is orthogonal, the differences between the marginal row and column means provide (best linear) unbiassed estimates of the differences between the corresponding parameters.

The formal similarity between this approach and ordinary regression techniques may be illustrated by the following device. A dummy variable taking only the values zero and unity is associated with each of the parameters in equation (1) which may be re-written as follows:

$$y_{ijk} = m\,v + a_i x_i + b_j z_j + e_{ijk} \qquad (8)$$

where $v = 1$ for all y_{ijk},

and $x_1 = 1$ for all y_{ijk} with $i = 1$, and $x_1 = 0$ when $i \neq 1$

$x_2 = 1$ for all y_{ijk} with $i = 2$, and $x_2 = 0$ when $i \neq 2$

. . . . . .

$x_r = 1$ for all y_{ijk} with $i = r$, and $x_r = 0$ when $i \neq r$

and $z_1 = 1$ for all y_{ijk} with $j = 1$, and $z_1 = 0$ with $j \neq 1$

. . . . . .

$z_s = 1$ for all y_{ijk} with $j = s$, and $z_s = 0$ when $j \neq s$

The moments matrix obtained from the dummy variables v,x and z by the application of least-squares estimation procedures is identical with the matrix of coefficients in equations (5). Obviously, sums of squares and cross-products for these dummy variables must be either zero or positive integers, and when they are not zero they have to be equal to the numbers of observations in particular cells, or combinations of cells, in the corresponding classification table. Thus, although in the non-orthogonal case it is necessary to solve a large block of simultaneous linear equations in order to obtain estimates of the parameters, the appropriate moments matrix or matrix of coefficients can be derived very much more easily and quickly than in a more conventional multivariate regression analysis.

The above method of analysis has obvious extensions to situations in which more than two sets of factors are assumed to influence the level of the dependent variable. For example, an extended model can be written as follows:

$$y_{ijklm} = m + a_i + b_j + c_k + d_l + e_{ijklm} \qquad (9)$$

If there is orthogonality in all four dimensions, the analysis will, of course, be relatively simple, but otherwise the computional difficulties are increased considerably as the number of parameters is increased.

If the underlying causal structure is adequately represented by a simple model of the types illustrated in equations (1) or (9), the analysis may be quite effective in explaining much of the variation in the dependent variable, since although there will generally tend to be a large number of parameters the number of possible different combinations of these parameters will be very large indeed. If, on the other hand, the effects of the various factors under consideration fail to combine in a simple additive fashion, this may lead to a a proliferation of interaction parameters which may make the analysis too clumsy both from a theortical and a computational viewpoint. Departures from additivity can be incorporated in the model by introducing a set of interaction parameters as in equation (10).

$$y_{ijk} = m + a_i + b_j + (ab)_{ij} + e_{ijk} \qquad (10)$$

Given that there are r parameters a_i and s parameters b_j, there are a possible rs interaction parameters of the type $(ab)_{ij}$, so that the number of simultaneous equations to be solved can increase from $(r + s)$ to $(r + s + rs)$ when the parameters are estimated by least-squares. Moreover, if there are k separate sets of factors, instead of just 2 as in equation (10), the number of possible sets of such first-order interactions is $\frac{k(k-1)}{2}$. It has been found necessary to introduce some interaction terms into the income analysis in this paper, but for obvious reasons the number has been kept as small as possible. For example, it is clear that the effects of age and occupation on income do not combine in a simple additive way; the earnings of manual workers tend to be highest for those workers who are in the middle of their working lives, whereas the earnings of professional workers tend to be highest just before retirement. There are fairly obvious reasons for this and any attempt to fit workers

from all occupations into the strait-jacket of a single set of age parameters will inevitably produce some sort of distorted compromise which is not representative of any group. Fortunately, however, it is not necessary to introduce a complete set of occupation-age interaction parameters since there is some justification for assuming that the effects of age on income are the same for certain groups of occupations even though they are not the same for all occupations. Accordingly, the following assumptions were made; the age pattern for manual workers, who comprise three-quarters of the adult-male labour force, was taken as the basic age pattern, and systematic departures from this pattern in other groups were allowed for by specifying one set of age interaction parameters for managerial and higher professional workers, another set for lower professional workers, and a third set for other non-manual workers. Since there are six age groups, this involves an additional 18 age-occupation parameters, but a complete set of interaction terms would have required 54 parameters as there are nine occupations altogether. Thus, in setting up the model it is assumed that certain age-occupation interaction parameters are shared among particular occupation, while others(for manual workers) are assumed to be zero. Apart from the extra computations involved the failure of the main effects to combine additively is a considerable loss in conceptual or theoretical simplicity, since it implies, for example, that occupational differentials are generally functions of age and vice versa. However, this problem arises from the complexity of the real world and not from any inadequacy of the method of analysis.

If the effects of the various factors combine in a multiplicative fashion, that is to say if the effect of any single factor is proportional to the existing level of the dependent variable, a simple transformation to the logarithm of the dependent variable is sufficient to restore additivity in the model. The multiplicative version of equation (10) is illustrated in equation (11) where the parameters require some

reinterpretation.

$$y_{ijk} = m\, a_i b_j (ab)_{ij} e_{ijk} \qquad (11)$$

The parameter m can be specified to be equal to the geometric mean of the y s, while the other parameters (and also the stochastic term) will generally take values around unity. If it assumed that the geometric mean of each set of parameters is unity the model is completely determinate. The stochastic terms are assumed to be always greater then zero with a geometric mean of unity. On taking logarithms, equation (11) can be rewritten as in equation (12)

$$\log y_{ijk} = \log m + \log a_i + \log b_j + \log (ab)_{ij} + \log e_{ijk} \qquad (12)$$

Thus, if the stochastic term e_{ijk} is log-normally distributed, the application of least-squares estimation procedures to equation (12) will also furnish maximum likelihood estimates of the parameters. When the dependent variable is income the assumption of log-normality for the residuals seems more plausible than that of normality, while the assumption of multiplicatively rather than additivity among the main effects also seems preferable.

It should also be noticed that in the multiplicative model the variance of the residuals is proportional to the expected value of income and is not homoscedastic except in logarithms. The association of dummy variables with the parameters of equation (11) can be accomplished by introducing then as exponents of the parameters. For example,

$$y_{ijk} = m^{v} a_i^{x_i} b_j^{z_j} (ab)_{ij}^{w_{ij}} e_{ijk} \qquad (13)$$

The dummy variables v, w, x and z are defined in a similar way to those in equation (8) above.

The moments matrix or matrix of coefficients of the independent variables which is obtained for the normal equations for equation (12) is exactly the same as for equation (10). If the equations are solved by matrix inversion, therefore, the additional computations involved in an investigation of the multiplicative model if a solution has already been obtained for the additive model are almost negligible (and vice versa). The two models can profitably be explored simultaneously in this case. Unfortunately, the writer did not have access to a high-speed computer when solving the equations for the income study in this paper, and solutions were obtained by an iterative method and not by matrix inversion. This also means that is impossible to provide confidence intervals for the various parameters or combinations of parameters in this study. The solution of such a large block of simultaneous equations by an iterative method without using high-speed computers is made feasible by the fact that the estimates of the parameters which are obtained by ignoring the non-orthogonality are generally quite good approximations to the proper estimates.*

---

* The method actually used was a systematic application of the method described by W.L. Stevens in [6].

III. Some Numerical Results

The specific model used to analyse the income data in this study is given in equation (13).

$$y_{ijklmn} = m + O_i + A_j + R_k + T_l + I_m + (OA)_{ij} + (OR)_{ik} + (OT)_{il} + e_{ijklmn} \quad (14$$

where O_i denotes occupation i (9 parameters)

A_j " age j (6 parameters)

R_k " region k (4 parameters)

T_l " town size l (3 parameters)

I_m " industry m (10 parameters)

$(OA)_{ij}$ " occupation-age interaction ij (18 parameters)

$(OR)_{ik}$ " occupation-region interaction ik (4 parameters)

$OT)_{il}$ " occupation-town size interaction il (3 parameters)

The following specifications are also added:

$$\Sigma n_i O_i = \Sigma n_j A_j = \ldots = \Sigma n_{il} (OT_{il}) = 0$$

where the n s denote the numbers of individuals affected by particular parameters. The corresponding equation for the multiplicative model is obtained simply by replacing the addition by multiplication signs (with appropriate reinterpretation of the parameters, stochastic term and side conditions). The numerical estimates obtained from the additive and multiplicative models are presented in Tables II and III below; the descriptions of the various sub-groups within each classification are listed in these tables together with their code numbers used for ease of reference here. Some comments are needed about the precise specification of the model, particularly in respect of the interaction terms. As already mentioned above, there are three sets of occupation-age interaction terms; the first set applies to occupations 1 and 2 only; the second set to occupation 3 only; and the third set to occupations 4,5 and 6 only. There are no interaction terms for occupations 7,8 and 9 (manual workers). The single

set of occupation-region interaction terms $(OR)_{ik}$ applies to occupations 1,2, and 3 only; these interaction terms allow for the possibility of the effects of location on income being different for managerial and profession workers from the effects for all other occupations. Similarly, there is a single set of occupation-town-size interaction parameters referring to occupations 1,2 and 3 only. Since there are likely to be interactions between industry and other factors, particularly occupation, it was decided to restrict the industry parameters to occupations 7 and 8 only (skilled and semi-skilled manual workers). The industry parameters are assumed to be zero for individuals in all other occupations. This is somewhat unsatisfactory but it avoids introducing numerous additional interaction parameters without ignoring the problem of interactions altogether. It was felt that the effects of industry on earnings are relatively most important for these two categories.

(See Tables II and III)

In Tables II and III the estimated parameters appear in the first column under the heading "fitted constant" to borrow the more expressive analysis of variance terminology. Instead of listing the interaction terms separately they have been added to the corresponding terms for the main effects since it is the sum of main effects plus interaction which yields the expected income pattern. For example, the expected age pattern within occupations 1 and 2 is derived by adding the interaction constants for these two occupations to the age constants common to all occupations. Four distinct age patterns are obtained in this way, three of which refer to the three occupational groups for which separate occupation-age interactions are postulated and one to the remaining group (manual workers) for which the interaction terms are assumed to be zero. In effect, the model specifies four age patterns and it is easier to present

the results in this form without listing the interaction terms separately although the latter can be derived by implication from the figures in the tables. The interactions between occupation and region and town size have been treated similarly.

The results for the multiplicative model shown in Table III are given as logarithms although strictly it is their anti-logarithms which denote the fitted constants; for example, the fitted constant for higher professional workers is 1.87, which is the anti-logarithm of 0.271, and this expresses the fact that earnings in this occupation (ignoring, for simplicity, the effects of all other factors) may be expected to be 87% above the (geometric) mean income (456) for the whole sample.

The actual deviations of the mean incomes for the various categories from the appropriate overall mean correspond to crude biassed estimates of the various parameters taking no account of the non-orthogonality of the distribution of the observations. These actual deviations are shown in the second column of Tables II and III for comparison with the fitted constants. The figures in the third column are the differences between these actual deviations and the fitted constants; these differences are labelled 'concentration components' since essentially they reflect how much of the actual deviation can be attributed to the effects of other factors through the concentration of particular types of observations in particular cells -- i.e., to the non-orthogonality of the data. The variability or dispersion of these concentration components relative to the dispersion of the corresponding fitted constants is a good indicator of the usefulness of the analysis. Whenever a specific category has a large individual concentration component it should be possible to adduce a specific explanation for it in terms of the actual distribution of observations. For example, the large positive concentration component for managers relative to higher professional workers is apparently due to the fact a much larger proportion of managers are found in the age groups 45-64 years where incomes are highest for these categories. The difference

between the average earnings of managers and higher professional workers therefore gives an exaggerated impression of the expected difference in earnings for individuals who are comparable in age and other respects. Again, it is clear that the difference in average earnings between workers in urban and in rural areas is at least partly due to differences in the occupational structure of the respective labour forces as well as possibly to other differences.

When there are serious interactions present in the data the interpretation of the so-called concentration components becomes much more difficult than when there are no interactions. The question of the effects of differences in age distribution among various occupations on the levels of income in those occupations is not altogether meaningful if the effects of age on income are not the same for all occupations. Thus, although the point of departure for this study was the problem of the non-orthogonality of the data, in practice the most serious difficulty is the presence of interactions. Indeed, the complexity of the present analysis and consequent proliferation of parameters and difficulties of interpretation suggest that this method would be most profitably applied to more restricted and homogeneous sets of data. Women, part-time workers and juveniles were deliberately excluded from the present study in an attempt to secure greater homogeneity, but even so the population examined seems to be too complex. The method of analysis might be much more powerful in a study of wage differentials where only manual workers are considered. For example, a study of relative levels of wages by industry must face the problem of differences in occupational structure among various industries, and the method of analysis described here might be ideally suited to take account of such differences. In this situation it would be much more plausible to assume additivity between the effects of occupation and industry on earnings, so that the problem resolves itself simply into disentangling the effects of the two factors when the distribution of the labour force is non-orthogonal. Age and location

could also easily be incorporated in such a study provided that the assumption of additivity is still considered satisfactory (and the appropriate data are available). Finally, it should be noticed that the problem of interactions is not peculiar to this method of analysis and is just as important in any conventional regression analysis. Indeed, the present analysis, as shown above, is essentially only a special case of regression analysis. The initial approach to the problem was more suggestive of variance analysis, but the latter can always be reduced to a form of regression analysis. The interactions show up clearly in the present study and it is essential to make explicit allowance for them, but the analogous difficulties in ordinary multivariate regression analysis are perhaps glossed over all too frequently.

Despite the clumsiness of the model presented above and the ensuing problems of computation and interpretation, a knowledge of the estimates of the various parameters enables much deeper insight to be gained into the way in which the factors considered influence the form of the distribution of incomes. Since specific allowance has been made for non-orthogonality in estimating the parameters, the effects of the various factors on income can be handled simultaneously and the following section will illustrate the uses to which the fitted constants may be put.

IV. Some Applications of the Numerical Results

The fitted constants estimated above can be combined in accordance with the specifications of the model to yield an expected level of income for each individual in the sample given his occupation, age, location, etc. A synthetic distribution of income based on these expected incomes can then be generated together with a corresponding distribution of residual (i.e., actual minus expected) incomes. These distributions may throw some light on the validity of certain hypotheses about the distribution of income.

The relative merits of the additive and the multiplicative versions of the model will be considered first. On the basis of Tables II and III there is little to choose between the two versions and both sets of estimated parameters exhibit the same basic pattern. Moreover, on the basis of the proportion of the variance of actual income (or the logarithm of income) which can be attributed to the variance of the expected level of income there is also little to choose between the two models. The relevant sums of squares are given in Table IV where a rough analysis of variance is presented. Strictly, the total of the sums of squares for the expected

(See Table IV)

and residual values of income should be identically equal to the sum of squares for actual income, but this relationship is only approximately satisfied by the data in Table IV. The main reason for the small discrepancies is that each sum of squares was calculated independently from the appropriate frequency distribution and there are grouping errors to be taken into account. It is clear from Table IV that both models succeed in explaining about a half of the variance of the dependent variable; this is analogous to obtaining a squared multiple correlation coefficient of about 0.5 in a multivariate regression analysis. In view of the large sample size, this suggests that we may reject the hypothesis that all the parameters are equal to each other, which, given the additional specifications of our model, is equivalent to the hypothesis that all the parameters are zero. A rough test of significance is carried out in Table IV; apart from the minor discrepancies in the sums of squares, there is the more serious problem that the observations are taken from a survey which departs in several respects from a simple random sample. There are grounds for believing that the sample obtained is less efficient from the present point of view than a simple random sample with the same number of observations. It can be argued, therefore, that a crude method of compensating for this would be to choose a smaller number of degrees of freedom for the whole sample than the actual number of observations. Fortunately, however, the F

ratios obtained in Table IV are so large that even if it were assumed that the appropriate number of degrees of freedom for the residual mean square was very much smaller than that shown in the table which is obtained on the assumption that the total number of degrees of freedom is one less than the total number of individuals (1,414), the null hypothesis would still have to be rejected. In any case, it is obvious from inspection of the fitted constants in Tables II and III that at least some of them, particularly some of the occupation and age constants, are significantly different from zero. It does not follow, of course, from this rough test that all the parameters are significantly different from zero, nor even that a particular set of parameters, such as those for region, are significantly different from zero and each other. In order to test the significance of a particular set of parameters it is necessary to recalculate all the fitted constants for a new model which omits these parameters, and then to derive a sum of squares for this particular set of parameters on the basis of the two models, i.e., the one including and the one excluding the parameters. This will generally be on exceedingly laborious operation since "the computations for these tests have little in common, so that one has almost to start from scratch with each hypothesis to be tested."* No attempt has therefore been

---

* See Oskar Kempthorne [4], p. 95.

---

made to test specific hypotheses of this type for the data analysed here. Of course, if standard errors were available for the individual parameters, or combinations of parameters, the situation would be much improved.

In Table V the synthetic distributions of expected incomes obtained from the additive and multiplicative versions of the model are compared with the distribution of actual income. First impressions of the synthetic distributions suggest a good

(See Table V)

superficial similarity to the typical positively skewed type of distribution commonly obtained from income data. Moreover, it is clear that it is by no means necessary to postulate that the effects of the various factors combine in a multiplicative rather than an additive fashion in order to generate a positively skewed distribution. Indeed, the additive model furnishes appreciably more individuals with expected incomes of over a £1,000 per year than the multiplicative model, the proportion for the additive model coinciding almost exactly with that actually observed for this range. The principal failure of both models is that they yield for too many individuals with expected incomes within the range £ 400 to £ 500 ; that is, the synthetic distributions are too peaked or leptokurtic compared with the actual distribution. Conversely, both models produce relatively too few individuals with expected incomes in the range £ 600 - £ 1000, and fail to produce any individuals with expected incomes below £ 200. However, the principal contrast between the two versions of the model is that whereas the additive model also has far too few individuals in the income group £ 300 - £ 400 and correspondingly too many in the group £ 500 - £ 600, the multiplicative model succeeds in producing approximately the correct proportions in these groups.

An objective test to decide which synthetic distribution resembles the actual distribution the more closely can be made by calculating a form of the statistic χ^2 for each of the synthetic distributions. The difference between the frequency in the synthetic distribution and the actual distribution in each of the class intervals in Table V is squared and divided by the actual frequency in the interval, and the results summed over all intervals. The statistics obtained are 15.4 and 12.1 for the additive and multiplicative cases respectively. This suggests that the latter distribution is a somewhat better approximation to the actual distribution than the former.

A second reason for preferring the multiplicative model emerges from a study of the residual incomes. The distribution of the residuals of actual income and logarithms of income are shown in Table VI. Both distribution are evidently unimodal

(See Table VI)

and fairly symmetrical, but tests for skewness suggest that there is still a little positive skewness in the residuals of actual income, although it is almost negligible in comparison with the skewness of the original distribution. The relevant statistics are given in Table VII where the skewness emerges clearly from measures based on the

(See Table VII)

relative positions of the mean and median, and first and ninth deciles, but not so clearly from the conventional measure using the quartiles. The residual logarithms display a barely perceptible negative skewness. However, despite the absence of any appreciable skewness in both distributions, neither distribution approximates very closely to a normal distribution since they both exhibit some leptokurtosis relatively to the normal distribution. This can also be seen from the statistics in Table VII where it is shown that, compared to a normal distribution of equal mean and variance, there are proportionately too many observations within the ranges the mean plus or minus one or two standard deviations and, conversely, far too many outside the ranges the mean plus or minus three or four standard deviations. Thus, in both distributions there are proportionately far too many extreme observations to conform to normality. This can also be shown by plotting the cumulative distributions on normal probability paper when S shaped curves are obtained.

Further examination reveals very marked heteroscedasticity in the residuals from the additive model; the variances of the residuals at different levels of expected income are given in Table VIII. While there is also some heteroscedasticity in the

(See Table VIII)

residuals of the logarithms of income, it is very much less than that for the residuals of actual income. It is interesting to see that the variance of the logarithmic residuals is actually smallest in the middle income ranges and not in the lowest income groups. The assumption that the variance of the residuals is proportional to the level of expected income is, of course, built into the multiplicative model, and this assumption is seen to be far more realistic than the assumption of the additive model that the residual variance is independent of expected income which is patently not true. On these grounds, therefore, the multiplicative model seems to be distinctly preferable to the additive model, and further analysis of the residuals will be confined to the logarithmic residuals.

It has frequently been argued that there are basic tendencies at work in the generation of the distribution of income which might cause it to be at least approximately log-normal.* The principal departure from normality for the distributions of both the

---

* See in particular [1], chapters 3 and 11.

---

logarithms of income and the residual logarithms of income considered here has already been shown to be leptokurtosis. However, it can be argued that better approximations to log-normality might still be found within particular sub-groups. There are two main lines of argument in favour of a log-normal distribution which will be briefly recapitulated here. The first is simply that the level of each individual's income is determined by the simultaneous operation of a large number of independent factors the effects of which combine together in a multiplicative fashion as illustrated in equation (15).

$$y^j = \prod_i^n e_i^j \quad \text{where} \quad e_i > 0 \ , \quad i = 1,2, \ldots, n \ . \qquad (15)$$

The supercript j identifies the jth individual. On taking logarithms, equation (16) can be written as follows:

$$\log\ y^j = \sum_i^n (\log\ e_i^j) \qquad (16)$$

According to the central limit theorem, if the terms e_i are mutually independent variables the distribution of log y will tend to normality for large n under fairly general conditions. Such a hypothesis might be advanced to explain the behaviour of the residuals in our multiplicative model even though it is clear from the above analysis that there are one or two important factors influencing income the effects of which are not mutually independent. The second line of argument reminds us of the fact that income is essentially a continuous flow through time and assumes that the proportionate changes in income from one period of time to the next may be represented by a stochastic variable. The basic assumption is expressed in equation (17).

$$\frac{y_t^j - y_{t-1}^j}{y_{t-1}^j} = \eta_t^j \qquad \eta_t > -1 \qquad (17)$$

where η_t^j is a random variable. Equation (17) may be reformulated as in equation (18):

$$y_t^j = y_{t-1}^j\ (1 + \eta_t) = y_{t-1}^j e_t^j \qquad e_t > 0 \qquad (18)$$

where e_t^j is similarly a random variable, which for simplicity may be assumed to have a geometric mean of unity (and a finite variance). The probability distribution of e_t^j is assumed to be the same for all individuals and invariant through time. Systematic

changes in income resulting from a trend or fluctuations in the general level of all incomes can easily be incorporated by introducing a coefficient c_t into equation (18) where c_t is assumed to be constant for all individuals in any one period but may vary from period to period. More generally, therefore, we may write:

$$y_t^j = c_t \, y_{t-1}^j \, e_t^j \qquad (19)$$

Each individual's income in period n may be expressed in terms of his initial income in some base period $t = 0$:

$$y_n^j = y_o^j \prod_{t=1}^{n} (c_t) \prod_{t=1}^{n} (e_t^j) \qquad (20)$$

By assumption, $\prod_{t=1}^{n} c_t$ is a constant for all individuals and may be replaced by the symbol C_n . On taking logarithms of equation (20) we have

$$\log \; y_n^j = \log \; y_o^j + \log \; C_n + \sum_{t=1}^{n} (\log \; e_t^j) \qquad (21)$$

If all individuals start from the same initial income y_o, the distribution of log y_n will be asymptotically normally distributed as n increases; the expected value of log y_n^j will be equal to the sum (log y_o + log C_n), while the variance of log y_n^j will be equal to the variance of log e multiplied by n, i.e., will increase linearly through time. Even if all individuals do not start from the same initial income the distribution of log y_n^j will still tend to normality provided n is large enough; in this case, it will evidently approach normality more quickly the larger the variance of log e_t relatively to the variance of log y_o, or, of course, if the initial distribution of y_o is itself not very different from log-normal. In any case, it can be deduced that the variance of log y_t will always increase by an

amount equal to the variance of log e between successive time periods.

There are certain implications of this line of argument to be noted. The first is that we expect to observe a log-normal distribution only for a group of individuals of roughly the same age who are also 'old' enough for the stochastic process outlined above to have generated a log-normal distribution. For, if we consider individuals of different ages, then we are in effect merging log-normal distributions with different variances (and possibly different means) so that the joint distribution will generally not be log-normal. One interesting possibility which arises in this sort of situation is that it can be shown that the result of pooling two or more normal distributions with a common mean but different variances is to produce a distribution which must exhibit leptokurtosis compared with a corresponding normal distribution.*

---

* See Appendix below for a proof of this relationship.

However, if the merged distributions also have different means, which may arise because the groups of individuals considered started from different average initial incomes or because they have different life cycles in income, then hardly any generalisation may be made about the form of the resulting joint distribution. Finally, it may be remarked that it does not follow from the fact that the variance of the logarithm of income will tend to increase through time for a particular group of individuals that the variance will increase for the population as whole when some sort of birth and death process for the members of the population is allowed for. It is quite feasible for the overall variance to be stable depending upon what assumptions are made about the proportion of new entrants each period and their initial income distribution, and also about the probabilities of extinction for existing members of the population of various ages.**

** See, for example, R.S.G. Rutherford [5], p. 280-283.

It should also be noticed that the variance (or indeed any conventional measure of dispersion) for a distribution of the logarithms of income is essentially a measure of the relative inequality of the distribution of income and not of the absolute variability of incomes. Clearly, the absolute dispersion of incomes will depend not simply on the variance of the logarithms but also on the general level of incomes. In the case of a log-normal distribution, for example, the variance of actual income is related to mean income and variance of the logarithms of income by the following equation: -

$$\text{var}_{y} = \bar{y}^2 (e^{\sigma^2} - 1)$$

where $\bar{y}$ is the arithmetic mean income and σ^2 is the variance of the logarithms of income.

The fact that the overall distributions of the logarithms of income and of the logarithmic residuals respectively are not normal does not preclude the possibility that normal distribution might occur within particular sub-groups. Indeed, the existence of differing normal distributions for particular sub-groups will be incompatible with a normal distribution for the entire population (and vice versa) except under special circumstances.* One subdivision of the population examined here which is immediately

---

* See [1] pp. 110, 111.

---

suggested is a distinction between salary and wage earners (roughly occupations 1 to 5 and 6 to 9 respectively in Tables II and III). The separate distributions for these two categories conform much more closely to log-normality than the overall distribution; this is true both of actual and of residual income, but especially the former. The measure of the improvement in approximation to normality is illustrated in Diagram I

(See Diagram I)

where the various cumulative logarithmic distributions are plotted on normal probability paper. It can be deduced from the statistics given in Table IX that

(See Table IX)

if the separate distributions for salary and wage earners were exactly log-normal, the joint distribution would not be log-normal given the differences in the variances of the logarithms of the two distributions. Even in the case of actual income where there is also some difference in the mean logarithms, the disparity in the variances is likely to be more important since it will greatly affect the tails of the merged distribution and manifest itself in form of leptokurtosis, which of course is actually what occurs. One interesting by-product of Table IX is that it shows that the model analysed above is relatively more successful in explaining the variance of the log-incomes of salary earners than that of wage earners, as indicated by the ratio of the residual to the actual variance. The reason for this is that age is a much more significant factor in explaining the variation of salary incomes than wage incomes. The heteroscedasticity of the residuals is also reflected by the data in Table IX, since the expected income for salary earners is, of course, higher than that for wage earners. Indeed, the gist of the present argument is that the principal reason why the distribution of the logarithmic residuals may not be normal is that the heteroscedasticity of the residuals is almost bound to result in leptokurtosis. If the distribution of the residuals were exactly normal at each level of expected income — a hypothesis which cannot be satisfactorilytested in a sample of this size — then heteroscedasticity would definitely imply leptokurtosis in the overall residual distribution compared with a corresponding normal distribution. The conjunction of heteroscedasticity and leptokurtosis of residuals is frequently observed in regression analyses based on cross-section data.

There is some positive correlation between age and the variance of the logarithms of actual income (see Table X). At first sight this seems to support the hypothesis

(See Table X)

outlined above which postulates a stochastic proportionate relationship between the level of income in successive time periods. However, it is also clear from the data in Table X that the reason for the increasing logarithmic variance is the fanning out of the life cycles in income for different occupational groups in the higher age groups. This is shown by the fact that the variance of the expected logarithmic incomes also increases with age, and indeed accounts for the whole of the increase in the variance of actual log incomes between age groups 2 and 6. The variance of the log residuals does not exhibit any relationship with age. The lowest age group is probably best ignored since it includes so many individuals in the course of training. It should also be noticed that when the sample is stratified into age groups or other categories, the sum of the residual and expected variances within each group does not have to be identical with the actual variance, although this relationship will tend to be approximately satisfied. The ratio of the residual variance to the actual variance declines uniformly with age, which mainly indicates that occupational differences are relatively more important in explaining variation in income in the higher than in the lower age groups.

The fact that it is systematic factors rather than the cumulative effect of some sort of stochastic process which accounts for the increasing variance of the logarithms of income is further illustrated by the second part of Table X. As soon as the sample is divided into salary and wage earners, it emerges that for wage earners, who account for four fifths of the sample, there is no obvious relationship between age and the variance of the logarithms of income. There is some tendency for this variance to increase with age for salary earners, at least over the age groups 3,4 and 5, but the

variance of residuals for salary earners turns out to be remarkably stable over age groups 2 to 6. Finally, it may be observed that in contrast to the subdivision of the sample into principal occupational groups, a subdivision by age does not appear to yield distributions which are closer approximations to the log normal distribution than the overall distribution. In particular, there seems to be no tendency for the distributions of the logarithms or their residuals to become more normal in the higher age groups, a tendency which is implied by the simple stochastic process hypothesis. A simultaneous classification into wage and salary earners and into two main age groups, namely 25 to 44 and 45 -64 years (the two small extreme age groups being excluded), produces four distributions each of which is approximately log-normal for the residuals and to a lesser extent for actual income. But neither for the residuals nor actual income does the additional breakdown by age reveal any improvement in approximation to log-normality, and indeed suggests some deterioration in the case of the actual income of salary earners. So far as the residuals are concerned, the distributions in the two age groups are virtually indistinguishable from each other for the whole sample, for wage earners separately and for salary earners separately, which might be inferred from the fact the variances of the residuals in Table X do not differ greatly between age groups (2) and (3) combined and (4) and (5) combined.

V. Conclusions

The principal problem in analysing the effects of factors such as occupation and age or income is not whether or not these effects combine in an additive or multiplicative fashion but rather that these factors combine or interact in a much more complicated way which cannot be approximated by either of these simple models. Allowance for these interactions greatly complicates the analysis but makes it possible to synthesize distributions of expected values of income which bear some resemblance to the actual empirical distribution. It is clear that it is not necessary to introduce

some form of "law of proportionate effect" in order to generate an income distribution which is markedly skew, and asymmetry is very largely eliminated in the distribution of residuals from the additive version of the model. However, the fact that the residual variation is an increasing function of the expected level of income is a strong argument in favour of postulating a law of proportionate effect, at least with respect to factors not specifically included in the model. Indeed, it appears that the residual variation may increase at a much faster rate than expected income as income rises. This in itself will tend to produce positive skewness in the distribution of actual income even if the distribution of expected incomes were perfectly symmetrical.

The particular model used was less successful in explaining the variation in manual workersearnings than the variation in salaries, despite the fact that a set of industry parameters was included specifically for skilled and semi-skilled workers. Since manual workers account for about 80 per cent of the sample, this relative failure accounts for the excessive bunching around the mode in the distribution of expected incomes. There seems to be more irrationality about the distribution of wages than salaries and it is probable that whereas a finer occupation classification might reduce the residual variation for salaried workers quite appreciably, it would not make much impression upon the distribution of wages. On these gounds it seems easier to explain the upper tail of the distribution where occupation and age are very powerful explanatory variables, than the central mass of the distribution. In general, it may be inferred that occupation is a much more significant explanatory variable in the higher age groups, while age is a much more significant explanatory variable among salaried workers.

If the labour force is divided into wage and salary earners, each employment income distribution conforms fairly well to a log-normal distribution. However, mainly

because of the difference in the variance of the logarithms of the two distributions the overall merged distribution does not, and indeed cannot, conform closely to a log-normal distribution. The same result is true of the residuals of the logarithms of income. Although no direct evidence in the form of positive correlation between the residual logarithmic variance and age was found to support a simple stochastic process model, the latter could perhaps be modified in such a way as not to imply an increasing variance through time. If some negative serial correlation were assumed between porportionate changes in income between successive time periods, the variance might be stabilised. A solution essentially along these lines has been suggested by Kalecki.* A restriction of this kind on the inherently explosive simple

---

* See [3].

---

stochastic proportionate relationship is intuitively plausible and not without indirect empirical support, since there is evidence to show that distributions of income on the basis or two or more years' income exhibit a smaller degree of inequality than distributions based on a single year's income. Finally, it has already been remarked that the merging of two or more approximately log-normal distributions, such as those for wage and salary earners, with differing variances will produce a distribution with proportionately too many extreme observations to conform to log-normality. If one distribution carries a much smaller weight than the other, as in this case, the effect of a greater variance in the less numerous distribution will only make itself felt in the tails of the merged distribution. That is, a reasonably good fit to a log-normal distribution will be observed over the range embracing the great majority of the observations in the pooled distribution, but the fit will be poor for the tails of the distribution. This is

generally the situation which is found when the upper tail of an income distribution conforms to a Pareto distribution, and it was pointed out at the outset of this paper that in our case the overall distribution suggests a good fit to a Pareto distribution for incomes above the mean. It seems quite feasible, therefore, that a reconciliation between the rival claims of a Pareto distribution and a log-normal distribution may be achieved by arguing that the latter is appropriate for more homogeneous sub-groups within the population and that the merging a series of log-normal distributions, which typically have variances inversely related to their sizes, tends in practice to yield overall distributions which conform approximately to a Pareto distribution.

Appendix

Leptokurtosis in merged normal distribution

Kurtosis is conventionally measured by the ratio of the fourth moment about the mean to the variance squared. In this proof it will be shown that the fourth moment of a distribution which is obtained by merging a series of normal distributions with identical means but differing variances must be greater than the fourth moment of a normal distribution with the same variance as the distribution obtained by the merging process.

Let v_i denote the variance of the ith normal distribution to be pooled to form a single merged distribution

Let w_i denote the weight or proportion of the merged distribution accounted for by the ith component normal distribution.

By definition $0 < w_i < 1$ and $\sum_i w_i \equiv 1$. Also, by assumption the mean of each component distribution is some constant. Let A refer to the merged distribution and B to a normal distribution of equal variance. Their common variance is given by: $\text{Var}^A = \text{Var}^B = \sum(w_i v_i)$ (1)

The fourth moments about the mean, μ_4^A and μ_4^B, may be written as follows by using the relationship $\frac{1}{3}\mu_4 \equiv \mu_2^2$ for normal distributions.

$$\tfrac{1}{3}\mu_4^A = \sum_i(w_i v_i^2); \qquad \tfrac{1}{3}\mu_4^B = \left\{\sum_i(w_i v_i)\right\}^2 \tag{2}$$

$$\therefore \tfrac{1}{3}\mu_4^B = \sum_i(w_i^2 v_i^2) + 2\sum_i\sum_j(w_i w_j v_i v_j) \qquad i \neq j \tag{3}$$

$$\therefore \tfrac{1}{3}\left(\mu_4^A - \mu_4^B\right) = \sum_i \left\{ w_i v_i^2 (1 - w_i) \right\} - 2 \sum_i \sum_j (w_i w_j v_i v_j) \qquad (4)$$

$$= \sum_i \left\{ w_i v_i^2 (\Sigma w_i - w_i) \right\} - 2 \sum_i \sum_j (w_i w_j v_i v_j)$$

$$= \sum_i \sum_j \left\{ w_i w_j (v_i^2 + v_j^2) \right\} - 2 \sum_i \sum_j (w_i w_j v_i v_j) \quad (i \neq j)$$

$$= \sum_i \sum_j \left\{ w_i w_j (v_i - v_j)^2 \right\} > 0 \qquad (5)$$

given $w_i > 0$ and $v_i \neq v_j$. This completes the proof.

TABLE I

Distribution of Wages or Salaries for Adult Male Employees Working Full-time throughout the Year (1953-54)

Income group (£ per annum)	Percentage of individuals
Less than 100	0.2
100 - 199	2.5
200 - 299	5.1
300 - 399	26.5
400 - 499	30.3
500 - 599	16.5
600 - 699	8.2
700 - 799	3.9
800 - 899	2.1
900 - 999	1.5
1,000 - 1,499	2.2
1,500 - 1,999	0.8
2,000 and over	0.3
Total	100
Mean	£ 496

Source: Oxford University Institute of Statistics, Survey of Incomes and Savings, 1954.

TABLE II

Comparison between Estimated Parameters or Fitted Constants and Actual Deviations about the Mean
(£ per annum)

	Occupational Group	Fitted constant	Actual deviation	Difference or concentration component
1.	Managerial	852	889	37
2.	Higher professional	452	452	0
3.	Lower professional	107	107	0
4.	Shopkeepers	102	109	7
5.	Clerical	14	-21	-35
6.	Service	-53	-30	23
7.	Skilled manual	-28	-30	- 2
8.	Semi-skilled manual	-89	-86	3
9.	Unskilled manual	-114	-122	- 8

	Age Group	Occupations 1 and 2 Fitted constant	Actual deviation	Concentration component	Occupation 3 Fitted constant	Actual deviation	Concentration component
1.	18-24	-688	-631	57	-271	-297	-26
2.	25-34	-146	-241	-95	- 87	-103	-16
3.	35-44	46	- 10	-56	36	38	2
4.	45-54	199	258	59	49	57	8
5.	55-64	434	489	55	157	178	21
6.	65 & over	-208	12	220	342	325	-17
		Occupations 4,5, and 6			Occupations 7,8 and 9		
1.	18-24	-203	-173	30	- 98	- 97	1
2.	25-34	- 6	- 20	-14	5	13	8
3.	35-44	130	130	0	43	35	- 8
4.	45-54	58	81	23	20	17	- 3
5.	55-64	- 66	- 80	-14	- 43	- 45	- 2
6.	65 & over	-129	- 81	48	- 20	- 76	-56

TABLE II (continued)

Comparison between Estimated Parameters or Fitted Constants and Actual Deviations about the Mean

(£ per annum)

	Sector	(Occupations 7 and 8 only) Fitted constant	Actual deviation	Concentration component
1.	Manufacturing: metal	19	28	9
2.	Manufacturing: other	3	12	9
3.	Agriculture	-57	- 125	-68
4.	Building & contracting	-45	-36	9
5.	Mining & quarrying	109	63	-46
6.	Transport & communications	-17	-11	6
7.	Public utilities	24	54	30
8.	Distribution	-68	-65	3
9.	Commerce & finance	-34	-42	- 8
10.	Public administration & defence	-11	9	20

	Region	Occupations 1,2 and 3 Fitted constant	Actual deviation	Concentration component	Occupations 4 to 9 Fitted constant	Actual deviation	Concentration component
1.	Scotland	-62	-218	-156	- 2	-11	- 9
2.	Northern England	-30	0	30	- 9	- 6	3
3.	Midlands & Wales	93	71	- 22	- 2	- 5	- 3
4.	Southern England	2	24	22	8	12	4
	Town Size						
1.	Conurbations	6	16	10	26	32	6
2.	Other urban	12	18	6	-18	-12	6
3.	Rural	-26	- 78	- 52	-24	-42	-18

TABLE III

Comparison between Estimated Parameters or Fitted Constants and Actual Deviations about the Mean
(LOGARITHMS of Income)

	Occupational Group	Fitted constant	Actual deviations	Difference or concentration component
1.	Managerial	.385	.406	.021
2.	Higher professional	.271	.271	.000
3.	Lower professional	.077	.081	.004
4.	Shopkeepers	.049	.057	.008
5.	Clerical	.031	- .007	- .038
6.	Service	- .045	- .020	.025
7.	Skilled manual	- .008	- .009	- .001
8.	Semi-skilled manual	- .065	- .063	.002
9.	Unskilled manual	- .087	- .091	- .004

		Occupations 1 and 2			Occupation 3		
	Age Group	Fitted constant	Actual deviation	Concentration component	Fitted constant	Actual deviation	Concentration component
1.	18-24	- .353	- .341	.012	- .257	- .282	- .025
2.	25-34	- .045	- .071	- .026	- .073	- .082	- .009
3.	35-44	.007	- .005	- .012	.046	.047	.001
4.	45-54	.096	.106	.010	.051	.058	.007
5.	55-64	.157	.176	.019	.118	.126	.008
6.	65 & over	- .012	.020	.032	.227	.218	- .009

		Occupations 4,5 and 6			Occupations 7,8 and 9		
1.	18-24	- .216	- .179	.037	- .122	- .123	- .001
2.	25-34	.011	- .002	- .013	.009	.015	.006
3.	35-44	.113	.119	.006	.045	.047	.002
4.	45-54	.054	.070	.016	.022	.020	- .002
5.	55-64	- .069	- .078	- .009	- .041	- .042	- .001
6.	65 & over	- .110	- .074	.036	- .018	- .072	- .054

TABLE III (continued)

Comparison between Estimated Parameters or Fitted Constants and Actual Deviations about the Mean
(LOGARITHMS of Income)

(Occupations 7 and 8 only)

	Sector	Fitted constant	Actual deviation	Concentration component
1.	Manufacturing: metal	.012	.019	.007
2.	Manufacturing: other	.009	.017	.008
3.	Agriculture	- .062	- .135	- .073
4.	Building & contracting	- .034	- .026	.008
5.	Mining & quarrying	.105	.058	- .047
6.	Transport & communications	- .011	- .003	.008
7.	Public utilities	.021	.053	.032
8.	Distribution	- .062	- .058	.004
9.	Commerce & finance	- .043	- .050	- .007
10.	Public administration & defence	- .005	.016	.021

		Occupations 1,2 and 3			Occupations 4 to 9		
	Region	Fitted constant	Actual deviation	Concentration component	Fitted constant	Actual deviation	Concentration component
1.	Scotland	- .027	- .136	- .109	- .008	- .020	- .012
2.	Northern England	- .027	- .015	.012	- .008	- .005	.003
3.	Midlands & Wales	.045	.039	- .006	- .004	- .007	- .003
4.	Sourthern England	.012	.031	.019	.011	.014	.003
	Town Size						
1.	Conurbations	.001	.008	.007	.027	.032	.005
2.	Other urban	.021	.025	.004	- .015	- .010	.005
3.	Rural	- .024	- .067	- .043	- .029	- .048	- .019

TABLE IV

Approximate Analysis of Variance

A. Actual Income (£ p.a.)

Factor	Sum of squares (millions)	Degrees of freedom	Mean square (thousands)	F ratio
Fitted constants	50.3	47	1070.2	30.1
Residuals	48.5	1366	35.5	
Total	101.6	1413	71.9	

B. Logarithms of Income

Factor	Sum of squares	Degrees of freedom	Mean square	F ratio
Fitted constants	19.8	47	0.4213	28.1
Residuals	20.5	1366	0.0150	
Total	41.0	1413	0.0290	

Note: The failure of the sums of squares to add exactly to the totals is explained in the text.

TABLE V

Income group	Actual income	Expected Income	
		Additive model	Multiplicative model
Less than 200	2.6	-	-
200 -	5.1	4.0	4.5
300 -	26.5	18.4	24.3
400 -	30.3	42.9	43.2
500 -	16.5	22.2	17.0
600 -	8.2	5.8	5.0
700 -	3.9	1.8	1.5
800 -	2.1	0.9	1.2
900 -	1.5	0.8	1.2
1,000 -	2.2	2.3	1.7
1,500 and over	1.1	0.9	0.4
Total	100	100	100

TABLE VI

Distribution of Income Residuals

Income Group £	Residual Income (additive model)	Income Group (logarithm of income)	Residual Income (multiplicative model)
-400 and less	1.1	-0.40 and less	0.9
-300 -	0.8	-0.30 -	0.3
-200 -	3.1	-0.20 -	2.7
-150 -	4.0	-0.15 -	4.3
-100 -	8.8	-0.10 -	6.3
- 50 -	16.9	-0.05 -	13.4
- 0 -	20.8	-0.00 -	21.5
+ 0 -	17.2	+0.00 -	19.9
+ 50 -	10.3	+0.05 -	14.5
+100 -	7.3	+0.10 -	8.3
+150 -	3.6	+0.15 -	3.9
+200 -	3.4	+0.20 -	3.1
+300 -	1.0	+0.30 -	0.7
+400 - and over	1.7	+0.40 and over	0.2
Total	100.0	Total	100.0

TABLE VII

Some Statistics Relating to the Distributions of Residuals

	Additive Model	Multiplicative Model
Mean	0 (£)	0.0000 (logarithms)
Median	-13 (£)	0.0012 (logarithms)
Standard deviation (σ)	185 (£)	0.1205 (logarithms)
Inter-quartile range	132 (£)	0.1369 (logarithms)
Skewness[1] (1)	0.053	-0.023
Skewness[2] (2)	0.101	-0.039
Skewness[3] (3)	0.206	-0.031

Percentage of observations lying outside the following ranges:

		(Normal curve)	
Mean $\pm\sigma$	29.04	(31.73)	22.83
Mean $\pm 2\sigma$	3.05	(4.55)	3.96
Mean $\pm 3\sigma$	1.32	(0.27)	1.49
Mean $\pm 4\sigma$	0.58	(0.006)	0.46

Notes: 1) First skewness measure = $\frac{(Q_3 - \text{Med}) - (\text{Med} - Q_1)}{(Q_3 - Q_1)}$ where Q_3 and Q_1 refer to upper and lower quartiles.

2) Second skewness measure = $\frac{(D_9 - \text{Med}) - (\text{Med} - D_1)}{(D_9 - D_1)}$ where D_9 and D_1 refer to ninth and first deciles.

3) Third skewness measure (Pearson's) using formula $\frac{3(\text{Mean} - \text{Med})}{\sigma}$

TABLE VIII

Expected Income (£)	Residual Income Variance	s.d.	Expected Log. Income	Residual Log. Income Variance	s.d.
200 - 399	3,700	61	2.3 - 2.5999	0.0149	0.122
400 - 499	13,300	115	2.6 - 2.6999	0.0125	0.112
500 - 699	21,300	146	2.7 - 2.7999	0.0123	0.111
700 - 999	68,500	262	2.8 - 2.9999	0.0260	0.161
1,000 and over	557,100	746	3.0 and over	0.0403	0.201

TABLE IX

Logarithms of Income

Category	Actual mean	Residual mean	Actual variance	Residual variance	Ratio of residual to actual variance
Salary earners	2.7811	-0.0019	0.0562	0.0239	0.425
Wage earners	2.6284	0.0002	0.0180	0.0122	0.678

TABLE X

Logarithms of Income

A. All Occupations

Age Group	Actual variance	Expected variance	Residual variance	Ratio of residual to actual variance
1. 18-24	0.0304	0.0046	0.0272	0.89
2. 25-34	0.0207	0.0046	0.0157	0.76
3. 35-44	0.0216	0.0082	0.0127	0.59
4. 45-54	0.0243	0.0111	0.0112	0.46
5. 55-64	0.0354	0.0236	0.0126	0.36
6. 65 and over	0.0406	0.0285	0.0129	0.32

B.	Salary Earners			Wage Earners		
Age Group	Actual variance	Residual variance	Ratio of residual to actual variance	Actual variance	Residual variance	Ratio of residual to actual variance
1. 18-24	0.0265	0.0196	0.74	0.3001	0.3001	1.00
2. 25-34	0.0404	0.0234	0.58	0.0142	0.0140	0.99
3. 35-44	0.0328	0.0260	0.80	0.0134	0.0088	0.66
4. 45-54	0.0446	0.0245	0.55	0.0133	0.0088	0.66
5. 55-64	0.0623	0.0240	0.39	0.0132	0.0100	0.76
6. 65 and over	0.0438	0.0233	0.53	0.0102	0.0087	0.85

References

[1] Aitchison, J. and J.A.C. Brown: "The Lognormal Distribution." Cambridge (England) 1957.

[2] Hill, T.P.: "Incomes, Savings and Net Worth; the Savings Surveys of 1952-54," The Bulletin of the Oxford University Institute of Statistics May 1955.

[3] Kalecki, M.; "On the Gibrat Distribution," Econometrica, April 1945.

[4] Kempthorne O.: "The Design and Analysis of Experiments." New York, 1952.

[5] Rutherford, R.S.G.: "Income Distributions: A New Model." Econometrica, July 1955.

[6] Stevens, W.L.: "Statistical Analysis of a Non-Orthogonal Tri-Factorial Experiment." Biometrika, Dec. 1948.