MISCELLANEA.

CONTENTS

THE ADJUSTMENT OF THE WEIGHTS OF COMPOUND INDEX NUMBERS BASED ON INACCURATE DATA.

By F. YATES, Sc.D.

WHEN it is desired to construct an index representing a complex entity, such as agricultural or industrial production for a series of years, or in a number of districts, it frequently happens that the component parts that go to make up this entity have been determined with very different accuracy. The question then arises, whether account should be taken of this variation in accuracy, and if so how.

It is assumed that the specification of the required entity is known. Thus agricultural production might be defined as the value of all agricultural produce ultimately available for human consumption, due allowance being made for imported feeding stuffs, produce consumed by the farmer himself, payments in kind, etc. The question of the need for specification is discussed by Kendall in a paper recently read before the Society, and published in the last number of the *Journal*, and by other speakers at the meeting.

While it is true that in the case of agricultural production data for the construction of such a quantitative measure are not at present widely available, sampling surveys of the type conducted by the Cambridge School of Agricultural Economics have shown that there is no insuperable difficulty in obtaining such data.

Any well-planned sampling survey should be so conducted that the sampling errors to which the various estimates are subject are capable of estimation from the results of the survey. To effect this it is necessary to introduce some element of randomization into the sampling. Such randomization, if properly planned and carried out, will automatically eliminate biases from the results.

Suppose that, armed with the results of such a survey, it is required to estimate for each of a number of districts, or for each of a number of years, an entity defined by

$$\Xi = \beta_1\xi_1 + \beta_2\xi_2 + \ldots + \beta_p\xi_p$$

where $\xi_1, \xi_2, \ldots \xi_p$ are quantities, varying from district to district or year to year, representing the true values of the component parts,

and $\beta_1, \beta_2, \ldots \beta_p$ are constants (price coefficients, etc.), whose values are determined by the definition of the entity. Instead of the function

$$X_0 = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

it is required to determine a function

$$X = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

where $x_1, x_2, \ldots x_p$ are the values of $\xi_1, \xi_2 \ldots \xi_p$ estimated from the survey, and $b_1, b_2, \ldots b_p$ are weighting constants to be determined, the function being such that differences between the different districts or years are represented as accurately as possible—*i.e.*, such that

$$V(X - \Xi) \text{ is minimum.}$$

If the conditions of random sampling are satisfied

$$x_1 = \xi_1 + e_1, x_2 = \xi_2 + e_2, \ldots$$

where $e_1, e_2, \ldots$ represent random sampling errors which are uncorrelated with $\xi_1, \xi_2, \ldots$

Let the variances and covariances of $\xi_1, \xi_2, \ldots$ be represented by

$$u_{11} = V(\xi_1), u_{12} = \text{Cov}(\xi_1 \xi_2), \ldots$$

and those of $e_1, e_2, \ldots$ by

$$v_{11} = V(e_1), v_{12} = \text{Cov}\,(e_1 e_2), \ldots$$

Actually, of course, the u's will be determined from the differences of the total variances and covariances of the x's over all districts or years and their sampling variances and covariances.

We have

$$V(X - \Xi) = (b_1 - \beta_1)^2 u_{11} + 2(b_1 - \beta_1)(b_2 - \beta_2)u_{12} + \ldots + b_1^2 v_{11} + 2b_1 b_2 v_{12} + \ldots$$

Differentiating with respect to $b_1, b_2, \ldots$ in turn, and equating to zero, we obtain the p linear equations

$$\begin{aligned} b_1 (u_{11} + v_{11}) + b_2 (u_{12} + v_{12}) + \ldots + b_p (u_{1p} + v_{1p}) &= \beta_1 u_{11} + \beta_2 u_{12} + \ldots + \beta_p u_{1p} \\ b_1 (u_{12} + v_{12}) + b_2 (u_{22} + v_{22}) + \ldots + b_p (u_{2p} + v_{2p}) &= \beta_1 u_{12} + \beta_2 u_{22} + \ldots + \beta_p u_{2p} \end{aligned}$$

These equations serve to determine the p coefficients $b_1, b_2, \ldots b_p$.

This procedure is analogous to that adopted in the formation of discriminant functions (Fisher, 1938), the final equations being identical with those obtained by Fairfield Smith (1936) in a somewhat similar problem.

The mean of the X's given by the above b's will not be equal to

the mean of the X_0's. It is therefore advisable to introduce the additional constant b_0 given by

$$b_0 = (\beta_1 - b_1)\,\bar{x}_1 + (\beta_2 - b_2)\,\bar{x}_2 + \ldots$$

If the interest centres in the actual values, rather than in the differences between the different districts or years, then it will be necessary to minimize

$$S(X - \Xi)^2.$$

The constant b_0 now enters directly into the equations, which, after simplification, reduce exactly to those already given.

The limitations of such an adjusted index must be clearly realized. Although it reduces the discrepancies between the true and the estimated values to a minimum, it inevitably introduces certain distortions in the process. In particular it may be misleading if we require to estimate the differences in production between a group of districts having a high value of some variate correlated with one or more of the ξ's and a group having a low value of the same variate. It is, in fact, always advisable to tabulate the values of X_0 given by the unadjusted weights as well as those of X.

Moreover, the variance of X is always less than the variance not only of X_0, but also of Ξ, being in fact

$$V(X) = b_1\beta_1 u_{11} + (b_1\beta_2 + b_2\beta_1)u_{12} + \ldots$$

At first sight this is perhaps startling, but in fact it provides the justification for a practice implicitly recognized by practical statisticians, but rarely explicitly stated—namely, that of reducing very high estimates and increasing very low ones when dealing with inaccurate material, as for example when making estimates of crop yields based on returns which are subject to considerable errors.

The analogy with partial regressions should also be noted. X is the partial regression function that would be obtained if the true values of the Ξ's were known and their regression on the observed x's determined.

It is a well-known property of regressions (though one which is occasionally overlooked) that when the independent variates are subject to error, the regression coefficients obtained by the ordinary methods will in general (subject to the inter-correlations existing between the independent variates) tend to be less than the true coefficients derived from the underlying physical law. A regression equation, in fact, gives the best prediction formula, but it does not, under these circumstances, give the best estimates of the true coefficients.

In the case of a single component, x_1, the equations reduce to

$$b_1 = \frac{u_{11}}{u_{11} + v_{11}}\,\beta_1$$

This is the equation giving the regression coefficient b_1 which will be obtained by an ordinary regression analysis when the regression on the true values of the independent variate (of which the determination is subject to error) is β_1.

References.

Fisher, R. A., "The Statistical Utilization of Multiple Measurements," *Ann. Eugenics*, 1938, VIII, 376–86.

Fairfield Smith, H., "A Discriminant Function for Plant Selection," *Ann. Eugenics*, 1936, VII, 240–50.