DEMOGRAPHY© Volume 15, Number 1 February 1978

A GENERAL METHOD OF DECOMPOSING A DIFFERENCE BETWEEN TWO RATES INTO SEVERAL COMPONENTS

Prithwis Das Gupta
Demography Unit, Indian Statistical Institute,
Calcutta 700 035, India

Abstract—In her work on the components of a difference between two rates, Kitagawa (1955) was successful in dividing the difference into the rate effect and the effect of the factor, for data classified by one factor. Her formulation for data classified by two factors, however, involves an interaction term which is difficult to interpret. Retherford and Cho (1973) devised a method that does not include any interaction terms. However, their method has other limitations, such as the dependence of the results on the order in which the effects of the factors are computed. In this paper, we provide a general method capable of handling any number of factors, which is developed along the lines suggested by Kitagawa and by Retherford and Cho but without the limitations of their methods.

INTRODUCTION

When interpreting the difference between two crude rates of the same phenomenon for two populations, demographers and other social scientists have always been cautious to recognize structural differences in the populations that might partly or wholly explain the difference. A comparison of two crude birth rates, for example, may differ from a comparison based on the corresponding "standardized" rates—standardizing, say, with respect to age and marital status. When calculating the difference between two rates, controlling for several factors, one might also want to study the contribution of each of these factors to the possible disparity between comparisons based on crude rates and standardized rates.

One of the earliest works dealing with this problem was by Goldfield (1948), whose method of multiple standardization, which allocated interactions equally to all factors involved in each interaction, was briefly described and used in John Durand's study of the U.S. labor force. This method was again used by Gibson (1975) in his study of the effects of changes in marital status and marital fertility on the decline in American fertility between 1961 and 1973. A comprehensive mathematical formulation of components analysis giving the relationships between the crude and the standardized rates for two groups of population was first presented by Kitagawa (1955). Her method was used by Blake and Das Gupta (1975) to study the motivational and technological components of the decline in American marital fertility between 1960 and 1970. Kitagawa primarily dealt with data cross-classified by two factors I and J; for data involving more factors, her suggestion was to combine them in some way so as to reduce them to two and then use the two-factor approach. Even in this two-factor case, her method involves, in addition to I and J effects, an interaction effect between the factors I and J, which cannot be easily interpreted.

In their study of recent fertility trends in East Asian countries, Retherford and Cho (1973) used a decomposition technique for two factors, I and J, that involved no in-

teraction term. This is surely an advantage over the Kitagawa method, but their method has other problems. First, the magnitudes of I and J effects depend on the order in which they are derived. Second, the method does not utilize the information about the factor J in the computation of I effect if I is the factor whose effect is derived first. Finally, although their rate (residual) effect can be interpreted as the difference between two conventional standardized rates (as in the case of Kitagawa's rate effect), it has two different expressions [as we will show later in equation (9)], depending on which of the factors I and J is considered first. On the other hand, the Kitagawa rate effect is unique, has a simpler expression [presented below in equation (3), second term] and poses no particular problems to merit rejection in favor of any other expression.

A general method of decomposition of a difference between two rates is presented in this paper along the lines suggested by Kitagawa and Retherford and Cho, but removing the limitations of their methods. More specifically, the present method (a) can be applied to data cross-classified by any number of factors, (b) gives a rate effect that is identical with that derived by the Kitagawa method, (c) does not involve any interaction terms, (d) gives results that are independent of the order in which the factors are considered, and (e) needs the entire cross-classified data for the computation of each and every effect.

In the last section, the general method is applied to study the change in the labor force participation rates in the United States between 1940 and 1970. Earlier applications of the present method can also be found in Blake and Das Gupta (1976) and Hernandez (1976) in their studies on the components of recent fertility decline in various countries.

DATA CLASSIFIED BY ONE FACTOR *I*

Let there be two populations called population 1 and population 2. These two populations can be, for example, the female population of the United States in the childbearing period in 1960 and 1970, or the population 14 years old and over in the states of California and Alabama (say, in 1970), or the total population of the United States and India (say, in 1975). For population 1, we use the following notations:

N_i = number of persons in the ith category of I,
E_i = number of events (such as births or deaths) in the ith category of I,
T_i = rate of persons in the ith category of I $(= E_i/N_i)$,
$N_.$ = total number of persons

$$\left(= \sum_i N_i\right),$$

$E_.$ = total number of events

$$\left(= \sum_i E_i\right),$$

and

$T_.$ = crude rate $(= E_./N_.)$.

For population 2, we use symbols analogous to N, E, and T—n, e, and t, respectively. If $N_i = 0$ (so that $E_i = 0$), we can set $T_i = t_i$. Similarly, if $n_i = 0$, we can set $t_i = T_i$.

The difference between the crude rates of populations 1 and 2 can be expressed as

$$t_. - T_. = \sum_i T_i\left(\frac{n_i}{n_.} - \frac{N_i}{N_.}\right) + \sum_i \frac{N_i}{N_.}(t_i - T_i) + \sum_i (t_i - T_i)\left(\frac{n_i}{n_.} - \frac{N_i}{N_.}\right). \quad (1)$$

The first term on the right-hand side of (1) measures the effect of changes in the I-composition, given the rates of population 1. The second term is the difference between two standardized rates with population 1 as standard. Unfortunately, these two terms do not add up to $(t_. - T_.)$, and we have a third term that accounts for the

interaction between rates and compositions.

In order to avoid the interaction term, Kitagawa suggests the alternative expression

$$t_{\cdot} - T_{\cdot} = \sum_i \frac{t_i + T_i}{2}\left(\frac{n_i}{n_{\cdot}} - \frac{N_i}{N_{\cdot}}\right) + \sum_i \frac{\frac{n_i}{n_{\cdot}} + \frac{N_i}{N_{\cdot}}}{2}(t_i - T_i), \quad (2)$$

where the first term on the right-hand side of (2) measures the effect of changes in the I-composition, given the average rates of populations 1 and 2, and the second term measures the effect of changes in the rates, given the average I-composition of populations 1 and 2.

We thus see that, as far as the one-factor case is concerned, Kitagawa was able to find a neat expression, equation (2), that decomposes the difference between two crude rates into I effect and rate effect. We should note here that if we distribute the third term on the right-hand side of equation (1) evenly between the first and the second terms, we obtain the two terms on the right-hand side of equation (2). Therefore, Kitagawa's formulation of the one-factor case is essentially the same as Goldfield's suggestion for the allocation of interactions. Kitagawa, however, poses the problem in a different way. She finds the effect of the change in the I-composition by assuming the *average* rates as standard, and also the effect of the change in the rates assuming the *average* I-composition as standard. It just so happens that these two effects add up to the difference between the two crude rates. Kitagawa's method, therefore, cannot be criticized for ignoring the interaction term in equation (1) or for distributing this term equally (and arbitrarily) between the effects.

DATA CLASSIFIED BY TWO FACTORS I AND J

For population 1, let us use the following notations:

N_{ij} = number of persons in the (i, j) category,

E_{ij} = number of events in the (i, j) category,

T_{ij} = rate of persons in the (i, j) category $(= E_{ij}/N_{ij})$,

$N_{i\cdot}$ = number of persons in the ith category of I

$$\left(= \sum_j N_{ij}\right),$$

$N_{\cdot j}$ = number of persons in the jth category of J

$$\left(= \sum_i N_{ij}\right),$$

$N_{\cdot\cdot}$ = total number of persons

$$\left(= \sum_i \sum_j N_{ij}\right),$$

$E_{i\cdot}$ = number of events in the ith category of I

$$\left(= \sum_j E_{ij}\right),$$

$E_{\cdot j}$ = number of events in the jth category of J

$$\left(= \sum_i E_{ij}\right),$$

$E_{\cdot\cdot}$ = total number of events

$$\left(= \sum_i \sum_j E_{ij}\right),$$

$T_{i\cdot}$ = rate of persons in the ith category of I $(= E_{i\cdot}/N_{i\cdot})$,

$T_{\cdot j}$ = rate of persons in the jth category of J $(= E_{\cdot j}/N_{\cdot j})$,

and

$T_{\cdot\cdot}$ = crude rate $(E_{\cdot\cdot}/N_{\cdot\cdot})$.

Analogous symbols are used for population 2 with n, e, and t substituted for N, E, and T, respectively. Thus, as in the one-factor case, if $N_{ij} = 0$, we set $T_{ij} = t_{ij}$; and if $n_{ij} = 0$, we set $t_{ij} = T_{ij}$.

The Kitagawa Method

Kitagawa first finds the combined *IJ* effect and the rate (residual) effect from equation (2) by considering the (i, j) categories as those of a single factor, as follows:

$$t_{..} - T_{..} = \sum_i \sum_j \frac{t_{ij} + T_{ij}}{2}\left(\frac{n_{ij}}{n_{..}} - \frac{N_{ij}}{N_{..}}\right) + \sum_i \sum_j \frac{\frac{n_{ij}}{n_{..}} + \frac{N_{ij}}{N_{..}}}{2} \cdot (t_{ij} - T_{ij}). \quad (3)$$

The first term on the right-hand side of equation (3), the combined *IJ* effect, is then further divided into *I* effect (independent of *J*), *J* effect (independent of *I*), and interaction (joint effect of *I* and *J*):

$$I \text{ effect} = \sum_i \sum_j \frac{t_{ij} + T_{ij}}{2} \frac{\frac{n_{.j}}{n_{..}} + \frac{N_{.j}}{N_{..}}}{2} \cdot \left(\frac{N_{ij}}{n_{.j}} - \frac{N_{ij}}{N_{.j}}\right), \quad (4)$$

$$J \text{ effect} = \sum_i \sum_j \frac{t_{ij} + T_{ij}}{2} \frac{\frac{n_{i.}}{n_{..}} + \frac{N_{i.}}{N_{..}}}{2} \cdot \left(\frac{N_{ij}}{n_{i.}} - \frac{N_{ij}}{N_{i.}}\right), \quad (5)$$

and

$$IJ \text{ interaction} = \sum_i \sum_j \frac{t_{ij} + T_{ij}}{2} \cdot \frac{\frac{N_{ij}n_{i.}}{N_{i.}n_{..}} - \frac{n_{ij}N_{i.}}{n_{i.}N_{..}} + \frac{N_{ij}n_{.j}}{N_{.j}n_{..}} - \frac{n_{ij}N_{.j}}{n_{.j}N_{..}}}{2}. \quad (6)$$

We, therefore, see that, unlike the one-factor case in equation (2), Kitagawa's formulation of the two-factor case does not decompose the difference between two crude rates into components without any interaction terms.

The Retherford-Cho Method

In order to do away with the interaction term, Retherford and Cho first find the *I* effect and the rate effect by ignoring the factor *J* and then applying the Kitagawa formula in equation (2) to the one-factor data. This gives

$$t_{..} - T_{..} = \sum_i \frac{t_{i.} + T_{i.}}{2}\left(\frac{n_{i.}}{n_{..}} - \frac{N_{i.}}{N_{..}}\right) + \sum_i \frac{\frac{n_{i.}}{n_{..}} + \frac{N_{i.}}{N_{..}}}{2}\left(t_{i.} - T_{i.}\right). \quad (7)$$

The second term on the right-hand side of equation (7), the provisional rate effect, is then further broken down into *J* effect and (final) rate effect by, again, applying the one-factor Kitagawa method in equation (2) to the differences $(t_{i.} - T_{i.})$:

$$J \text{ effect} = \sum_i \sum_j \frac{t_{ij} + T_{ij}}{2} \cdot \frac{\frac{n_{i.}}{n_{..}} + \frac{N_{i.}}{N_{..}}}{2}\left(\frac{n_{ij}}{n_{i.}} - \frac{N_{ij}}{N_{i.}}\right), \quad (8)$$

and

$$\text{Rate effect} = \sum_i \sum_j \frac{\frac{n_{i.}}{n_{..}} + \frac{N_{i}}{N_{..}}}{2} \cdot \frac{\frac{n_{ij}}{n_{i.}} + \frac{N_{ij}}{N_{i.}}}{2}(t_{ij} - T_{ij}). \quad (9)$$

We notice from a comparison of (5) and (8) that the *J* effects in the Kitagawa method and the Retherford-Cho method are identical. Although the decomposition in equations (7)–(9) does not include any interaction terms, it is evident that (a) the rate effect in equation (9) in the Retherford-Cho method is not identical with the rate effect in equation (3) obtained by Kitagawa, and the latter should be accepted for reasons discussed earlier; (b) the *I* effect in (7) is independent of the way the factor *J* is distributed within the categories of the factor *I*; and (c) the *I* and *J* effects would not be the same as in equa-

tions (7) and (8) if the J effect were computed first.

An Alternative Approach

Our starting point is the correct decomposition of the difference between two crude rates into a combined IJ effect and a rate effect, as done by Kitagawa in equation (3). In order to further decompose the combined IJ effect in (3) into I and J effects, we visualize two populations, 1′ and 2′, in which the population sizes in the (i, j) categories are the same as in populations 1 and 2, respectively, but, in both of them, the rate of persons in the (i, j) category is $(t_{ij} + T_{ij})/2$. Denoting by R and r the rates in populations 1′ and 2′, we, therefore, have

$$R_{ij} = r_{ij} = \frac{r_{ij} + R_{ij}}{2} = \frac{t_{ij} + T_{ij}}{2}, \qquad (10)$$

$$\frac{r_{i.} + R_{i.}}{2} = \frac{1}{2}\left(\sum_j \frac{r_{ij}n_{ij}}{n_{i.}} + \sum_j \frac{R_{ij}N_{ij}}{N_{i.}}\right)$$

$$= \sum_j \frac{t_{ij} + T_{ij}}{2} \frac{\frac{n_{ij}}{n_{i.}} + \frac{N_{ij}}{N_{i.}}}{2}, \qquad (11)$$

$$\frac{r_{.j} + R_{.j}}{2} = \sum_i \frac{t_{ij} + T_{ij}}{2} \frac{\frac{n_{ij}}{n_{.j}} + \frac{N_{ij}}{N_{.j}}}{2}, \qquad (12)$$

and

$$r_{..} - R_{..} = \text{the same as the combined } IJ \text{ effect in (3).} \qquad (13)$$

Let us now apply the Retherford-Cho method in equations (7)–(9) to decompose $(r_{..} - R_{..})$ for populations 1′ and 2′ into I, J, and rate effects. We note that, because of (10), the rate effect will vanish. Also, because of (13), this decomposition will be a decomposition of the combined IJ effect in populations 1 and 2 into I and J effects. Again, since equations (7)–(9) can be applied either in the order (I, J) or (J, I), let us denote by $I(I, J)$ and $J(I, J)$ the I and J effects when derived in the order I and J, and by $I(J, I)$ and $J(J, I)$ the I and J effects when the J effect is obtained first. From equations (7) and (11), we obtain

$$I(I, J) = \sum_i \sum_j \frac{t_{ij} + T_{ij}}{2} \frac{\frac{n_{ij}}{n_{i.}} + \frac{N_{ij}}{N_{i.}}}{2} \cdot \left(\frac{n_{i.}}{n_{..}} - \frac{N_{i.}}{N_{..}}\right), \qquad (14)$$

and (8) and (10) give

$$J(I, J) = \sum_i \sum_j \frac{t_{ij} + T_{ij}}{2} \cdot \frac{\frac{n_{i.}}{n_{..}} + \frac{N_{i.}}{N_{..}}}{2} \left(\frac{n_{ij}}{n_{i.}} - \frac{N_{ij}}{N_{i.}}\right). \qquad (15)$$

The J effect in (8) and $J(I, J)$ in (15) are obviously identical. From symmetry, $J(J, I)$ and $I(J, I)$ have expressions analogous to $I(I, J)$ and $J(I, J)$, respectively. The only changes necessary are the replacements of $n_{i.}$ and $N_{i.}$ by $n_{.j}$ and $N_{.j}$, respectively.

Since it is as logical to compute first the J effect as it is to compute the I effect, we finally define the I and J effects as the arithmetic means of the corresponding values obtained from the two possible orders. In other words,

$$I \text{ effect} = \tfrac{1}{2}[I(I, J) + I(J, I)],$$

$$J \text{ effect} = \tfrac{1}{2}[J(I, J) + J(J, I)]. \qquad (16)$$

The I and J effects in (16), together with the rate effect equal to the second term on the right-hand side of (3), constitute the proposed method of decomposition of the difference between two crude rates for data classified by two factors.

The Relationship Between the Kitagawa and the Present Formulations

We notice that the J effect in equation (5) and $J(I, J)$ in (15) are identical. From symmetry, the I effect in (4) is also identical with $I(J, I)$. Therefore, Kitagawa's decomposition of the combined IJ effect (say, C) is

$$C = I(J, I) + J(I, J) + IJ \text{ interaction.} \quad (17)$$

Also, from (14) and (15),

$$C = I(I,J) + J(I,J) = I(J,I) + J(J,I). \quad (18)$$

Therefore, from (16) and (18), and then, from (17), we obtain

$$I \text{ effect} = \tfrac{1}{2}[\{C - J(I, J)\} + I(J, I)]$$

$$= \tfrac{1}{2}[\{I(J, I) + IJ \text{ int.}\} + I(J, I)]$$

$$= I(J, I) + \frac{IJ \text{ int.}}{2}. \quad (19)$$

Similar arguments also lead to

$$J \text{ effect} = J(I, J) + \frac{IJ \text{ int.}}{2}. \quad (20)$$

It is evident from a comparison of (17) with (19) and (20) that, if Kitagawa's IJ interaction term in (6) is evenly allocated to her I and J effects in (4) and (5), we obtain the I and J effects in equation (16) by the present method. This is again consistent with Goldfield's approach and also with the Kitagawa formulation in (2) for data classified by one factor. Here again it may be mentioned that the idea behind the present approach (like the Kitagawa one-factor approach) is to find the effect of the change in one factor, holding the other factors constant at an *average* level, and, fortunately, all such effects add up to the difference between the two crude rates. Therefore, the question of why Kitagawa's interaction term in (6) was distributed evenly between the factors, and not in some other proportions, does not seem to be a valid one.

DATA CLASSIFIED BY THREE FACTORS I, J, AND K

Because of complexity of expressions, Kitagawa suggests that, when there are three factors I, J, and K, one of them, say I, may be considered as one factor and the cross-classification of J by K as a second factor. This will enable one to use Kitagawa's two-factor formulas in a three-factor situation. The alternative method suggested in the preceding section for two factors can, however, be extended directly to three factors. Using analogous notations, we first divide the difference between the two crude rates ($t_{...} - T_{...}$) into the following components, as done by Kitagawa:

$$\text{Combined } IJK \text{ effect} = \sum_i \sum_j \sum_k \frac{t_{ijk} + T_{ijk}}{2}\left(\frac{n_{ijk}}{n_{...}} - \frac{N_{ijk}}{N_{...}}\right), \quad (21)$$

$$\text{Rate effect} = \sum_i \sum_j \sum_k \frac{\frac{n_{ijk}}{n_{...}} + \frac{N_{ijk}}{N_{...}}}{2}(t_{ijk} - T_{ijk}). \quad (22)$$

Again using similar notations for the effects for a particular order of I, J, and K, we find expressions analogous to (14) and (15) as follows:

$$I(I, J, K) = \sum_i \sum_j \sum_k \frac{t_{ijk} + T_{ijk}}{2} \cdot \frac{\frac{n_{ijk}}{n_{i..}} + \frac{N_{ijk}}{N_{i..}}}{2}\left(\frac{n_{i..}}{n_{...}} - \frac{N_{i..}}{N_{...}}\right), \quad (23)$$

$$J(I, J, K) = \sum_i \sum_j \sum_k \frac{t_{ijk} + T_{ijk}}{2} \cdot \frac{\frac{n_{ijk}}{n_{ij.}} + \frac{N_{ijk}}{N_{ij.}}}{2} \cdot \frac{\frac{n_{i..}}{n_{...}} + \frac{N_{i..}}{N_{...}}}{2}\left(\frac{n_{ij.}}{n_{i..}} - \frac{N_{ij.}}{N_{i..}}\right), \quad (24)$$

and

$$K(I, J, K) = \sum_i \sum_j \sum_k \frac{t_{ijk} + T_{ijk}}{2} \cdot \frac{\frac{n_{i..}}{n_{...}} + \frac{N_{i..}}{N_{...}}}{2} \cdot \frac{\frac{n_{ij.}}{n_{i..}} + \frac{N_{ij.}}{N_{i..}}}{2}\left(\frac{n_{ijk}}{n_{ij.}} - \frac{N_{ijk}}{N_{ij.}}\right). \quad (25)$$

The expressions for the I, J, and K effects corresponding to other orders in which they can be computed follow directly from equations (23)–(25) by proper substitution of suffixes. The effect $K(J, K, I)$, for example, has the same expression as that for $J(I, J, K)$ in (24), except that $n_{ij.}$, $N_{ij.}$, $n_{i..}$, and $N_{i..}$ have to be replaced by $n_{.jk}$, $N_{.jk}$, $n_{.j.}$, and $N_{.j.}$, respectively. We note that $I(I,J,K,) = I(I,K,J)$, $J(J,I,K) = J(J,K,I)$, and $K(K,I,J) = K(K,J,I)$. As in (16), we finally obtain the I, J, and K effects as

$$I \text{ effect} = \tfrac{1}{6}[I(I,J,K) + I(I,K,J) + I(J,I,K) + I(J,K,I) + I(K,I,J) + I(K,J,I)], \quad (26)$$

$$J \text{ effect} = \tfrac{1}{6}[J(I,J,K) + J(I,K,J) + J(J,I,K) + J(J,K,I) + J(K,I,J) + J(K,J,I)], \quad (27)$$

and

$$K \text{ effect} = \tfrac{1}{6}[K(I,J,K) + K(I,K,J) + K(J,I,K) + K(J,K,I) + K(K,I,J) + K(K,J,I)]. \quad (28)$$

DATA CLASSIFIED BY FOUR FACTORS $I, J, K,$ AND L

When four factors are involved, the difference $(t_{....} - T_{....})$ is first broken down into two components, as in equations (21)–(22), as follows:

Combined $IJKL$ effect =

$$\sum_i \sum_j \sum_k \sum_l \frac{t_{ijkl} + T_{ijkl}}{2}\left(\frac{n_{ijkl}}{n_{....}} - \frac{N_{ijkl}}{N_{....}}\right), \quad (29)$$

Rate effect =

$$\sum_i \sum_j \sum_k \sum_l \frac{\frac{n_{ijkl}}{n_{....}} + \frac{N_{ijkl}}{N_{....}}}{2}(t_{ijkl} - T_{ijkl}). \quad (30)$$

The combined $IJKL$ effect in (29) is then further divided into the effects of four individual factors I, J, K, and L:

$$I \text{ effect} = \tfrac{1}{24}[I(I,J,K,L) + I(I,J,L,K) + \cdots + I(L,K,I,J) + I(L,K,J,I)], \quad (31)$$

$$J \text{ effect} = \tfrac{1}{24}[J(I,J,K,L) + J(I,J,L,K) + \cdots + J(L,K,I,J) + J(L,K,J,I)], \quad (32)$$

$$K \text{ effect} = \tfrac{1}{24}[K(I,J,K,L) + K(I,J,L,K) + \cdots + K(L,K,I,J) + K(L,K,J,I)], \quad (33)$$

and

$$L \text{ effect} = \tfrac{1}{24}[L(I,J,K,L) + L(I,J,L,K) + \cdots + L(L,K,I,J) + L(L,K,J,I)]. \quad (34)$$

Each of the above effects is the average of 24 such effects, corresponding to 24 (= 4!) possible orders in which the effects of four factors can be computed. The effects for any particular order can be obtained from the following typical expressions, similar to (23)–(25):

$$I(I,J,K,L) = \sum_i \sum_j \sum_k \sum_l \frac{t_{ijkl} + T_{ijkl}}{2} \cdot \frac{\frac{n_{ijkl}}{n_{i...}} + \frac{N_{ijkl}}{N_{i...}}}{2}\left(\frac{n_{i...}}{n_{....}} - \frac{N_{i...}}{N_{....}}\right), \quad (35)$$

$$J(I,J,K,L) = \sum_i \sum_j \sum_k \sum_l \frac{t_{ijkl} + T_{ijkl}}{2} \cdot \frac{\frac{n_{ijkl}}{n_{i...}} + \frac{N_{ijkl}}{N_{ij..}}}{2} \cdot \frac{\frac{n_{i...}}{n_{....}} + \frac{N_{i...}}{N_{....}}}{2}\left(\frac{n_{ij..}}{n_{i...}} - \frac{N_{ij..}}{N_{i...}}\right), \quad (36)$$

$$K(I, J, K, L) =$$

$$\sum_i \sum_j \sum_k \sum_l \frac{t_{ijkl} + T_{ijkl}}{2} \frac{\frac{n_{ijkl}}{n_{ijk.}} + \frac{N_{ijkl}}{N_{ijk.}}}{2}$$

$$\cdot \frac{\frac{n_{i...}}{n_{....}} + \frac{N_{i...}}{N_{....}}}{2} \frac{\frac{n_{ij..}}{n_{i...}} + \frac{N_{ij..}}{N_{i...}}}{2} \left(\frac{n_{ijk.}}{n_{ij..}} - \frac{N_{ijk.}}{N_{ij..}}\right), \quad (37)$$

$$L(I, J, K, L) =$$

$$\sum_i \sum_j \sum_k \sum_l \frac{t_{ijkl} + T_{ijkl}}{2}$$

$$\cdot \frac{\frac{n_{i...}}{n_{....}} + \frac{N_{i...}}{N_{....}}}{2} \frac{\frac{n_{ij..}}{n_{i...}} + \frac{N_{ij..}}{N_{i...}}}{2}$$

$$\cdot \frac{\frac{n_{ijk.}}{n_{ij..}} + \frac{N_{ijk.}}{N_{ij..}}}{2} \left(\frac{n_{ijkl}}{n_{ijk.}} - \frac{N_{ijkl}}{N_{ijk.}}\right). \quad (38)$$

If the numbers of categories of the factors I, J, K, and L are, respectively, m_i, m_j, m_k, and m_l, and if any of these numbers, say, m_l is equal to 1, then the data involving four factors I, J, K, and L degenerate into a three-factor case involving the factors I, J, and K. In such a situation, symbols such as $N_{ijk.}$, N_{ijkl}, or N_{ijk} would refer to the same number. Keeping this in mind, if we substitute $m_l = 1$ in the four-factor formulas (29)–(38), we obtain the formulas (21)–(28) corresponding to the three-factor case. Similarly, substitution of $m_k = m_l = 1$ and $m_k = m_l = m_j = 1$ in the four-factor formulas (29)–(38) gives, respectively, the two-factor formulas (14)–(16) and the one-factor formula (2) by Kitagawa. This is another reason, besides the one mentioned earlier (p. 104), why the present method is consistent with, and can be regarded as a generalization of, Kitagawa's approach for one-factor data. Also, the fact—that a simple substitution of 1 for one or more of the m values in the formulas renders the formulas equivalent to those for lower order cross-classified data—makes it possible to write a general computer program for the components analysis of, say, four-factor data, and use it for any set of data involving one, two, three, or four factors (see next section).

In most situations, because of nonavailability of data classified by more than four factors, or because of insufficient numbers in the cells even if such data were available, consideration of formulas appropriate for five or more factors would largely be a matter of academic interest. However, if the situation demands, it is obvious from the symmetry of the four-factor formulas (29)–(38) how to write down the rate effect and the effects of factors when any number of factors is involved. As more and more factors are introduced, the computations become increasingly lengthy. It was indeed a formidable task to consider even three factors when Kitagawa's article appeared in 1955. Because of the advent of electronic computer facilities since then, computation for a reasonable number of factors is no longer a problem, once we have been able to systematically develop a general method within a tight framework.

A NUMERICAL ILLUSTRATION

We now illustrate the application of the present method with U.S. data for 1940 and 1970. The percent of population (14 years old and older) in the labor force was 52.22 in 1940 (population 1) and 55.49 in 1970 (population 2), so that the difference between the two crude labor force participation rates is 3.27. For both years, data are available by cross-classifying the total population aged 14 years old and older (persons) and the population in the labor force (events) by age, sex, marital status, and region (Appendix Tables 1–4). The categories considered for these factors are as follows:

I (Age): 14 to 24, 25 to 34, 35 to 44, 45+
L (Sex): 1 = male, 2 = female
K (Marital Status): 1 = single, 2 = married–spouse present, 3 = other
J (Region): 1 = urban, 2 = rural-nonfarm, 3 = rural-farm.

With the above available data and their marginal totals, it is possible to do four one-factor, six two-factor, four three-factor, and one four-factor components analysis of the difference, 3.27, between the two crude rates. A computer program was written for a four-factor analysis, but, if data cards are provided according to the simple instructions given in the comment cards of the program, the program can as well handle one-, two-, or three-factor cases (the program is available from the author on request). As a matter of fact, the program was run only once with all the fifteen possible sets of data put together, and the summary results are presented in Table 1. We have shown all these results only for the purpose of illustration. In reality, however, we will be satisfied with only those results that correspond to the maximum utilization of data (the four-factor results in Table 1 in the present case). For the latter case, the age-sex-marital status-region-standardized rates are 51.39 and 55.81 for 1940 and 1970, respectively, and their difference (4.42) serves as the rate effect. It may be worth mentioning that it took less than 1.5 seconds for a 7600 CDC computer to do all the computations with fifteen sets of data.

Table 1.—Components Analysis of the Difference Between the Crude Labor Force Participation Rates (per 100) for Persons Aged 14 Years and Older: United States, 1940 and 1970

Number of Factors	Effects of Factors				Rate Effect	Total Effect
	Age	Sex	Marital Status	Region		
1	-1.28				4.55	3.27
		-0.94			4.21	3.27
			0.06		3.21	3.27
				0.85	2.42	3.27
2	-1.33	-0.98			5.58	3.27
	-1.48		-0.16		4.91	3.27
	-1.34			0.82	3.79	3.27
		-0.61	0.38		3.50	3.27
		-0.78		1.01	3.04	3.27
			0.12	0.94	2.21	3.27
3	-1.51	-0.63	0.19		5.22	3.27
	-1.41	-0.84		0.96	4.56	3.27
	-1.56		-0.11	0.82	4.12	3.27
		-0.54	0.37	1.01	2.43	3.27
4	-1.62	-0.58	0.16	0.89	4.42	3.27

Sources: U.S. Bureau of the Census, *Census of the United States: 1940, Population*, Vol. 3, The Labor Force, Part 1: United States Summary (Washington, D.C.: U.S. Government Printing Office, 1943), Table 6; U.S. Bureau of the Census, *Census of the United States: 1940, Population*, Vol. 4, Characteristics by Age, Part 1; United States Summary (Washington, D.C.: U.S. Government Printing Office, 1943), Tables 6, 8, and 9; U.S. Bureau of the Census, *Census of Population: 1970, Detailed Characteristics*, Final Report PC(1)-D1, United States Summary (Washington, D.C.: U.S. Government Printing Office, 1973), Tables 203, 215, and 216.

Regarding the results in Table 1, we notice that the effects of the factors vary within a narrow range depending on how many and what factors are considered. On the other hand, the rate effect shows a considerable amount of fluctuation, having a relatively high or low value depending, respectively, on whether the factors with negative or positive effects predominate. Based on the four-factor analysis in Table 1, we present our major conclusions in a tabular form in Chart 1 (where, by rate, we mean an *IJKL*-specific rate).

Chart 1

If the following components were the same (and equal to the average of 1940 and 1970) in 1940 and 1970	And only the following component(s) changed as it (they) did from 1940 to 1970	The change in percent of population in labor force from 1940 to 1970 would be
Sex, Marital Status, Region, Rate	Age	a *decrease* of 1.62
Age, Marital Status, Age, Sex, Region, Rate	Sex Marital Status	a *decrease* of 0.58 an *increase* of 0.16
Age, Sex, Marital Status, Rate	Region	an *increase* of 0.89
Age, Sex, Marital Status, Region	Rate	an *increase* of 4.42
	Age, Sex, Marital Status, Region, Rate	an *increase* of 3.27

Two general remarks made by Kitagawa about the components analysis are worth repeating. First, the effects of factors do not necessarily imply any causal relationships. They simply indicate the nature of the association of the factors with the phenomenon being measured. There might be some hidden forces behind the factors that are actually responsible for the numbers we allocate to different factors as effects, but identifying these forces is beyond the scope of the components analysis. Second, the total effect is not necessarily explained more and more by increasing the number of factors. In other words, we do not expect the rate effect to gradually decrease with an increase in the number of factors (see Table 1). The difference between two standardized rates is not necessarily less than the difference between the two corresponding crude rates, whereas, the residual variance in a regression problem is necessarily less than the total variance. Components analysis, therefore, is not like regression analysis where the addition of each independent variable to the equation increasingly explains the variation in the dependent variable. Another major difference between components analysis and regression analysis is that, given the data, the criterion that separates the two populations in a components analysis may itself be treated as an independent variable in a regression analysis. If, for example, data are available on the IQ's of children and also on their race (white/negro), their family's income, and the education of their mother, a components analysis may be used to attempt to find how much of the average 15-point difference between the IQ's of white and negro children can be explained in terms of differences in family income and education. On the other hand, a regression analysis may address itself to the

question of how much of the variation in the IQ's of the children can be explained by race, family income, and education.

REFERENCES

Blake, Judith, and P. Das Gupta. 1975. Reproductive Motivation Versus Contraceptive Technology: Is Recent American Experience an Exception? Population and Development Review 1:229-249.

———, and P. Das Gupta. 1976. Components of the Decline in American Marital Fertility Between 1960 and 1970. Manuscript. Berkeley: University of California.

Gibson, Campbell. 1975. Changes in Marital Status and Marital Fertility and Their Contribution to the Decline in Period Fertility in the United States: 1961-1973. Paper presented at the annual meeting of the Population Association of America, April 1975, in Seattle, Washington.

Goldfield, E. D. 1948. Appendix B: Methods of Analyzing Factors of Labor Force Change. Pp. 219-236 in John D. Durand, The Labor Force in the United States: 1890-1960. New York: Social Science Research Council.

Hernandez, Donald J. 1976. Policy Vs. Other Factors in Fertility Decline After 1950. Unpublished Ph.D dissertation. Berkeley: University of California.

Kitagawa, E. M. 1955. Components of a Difference Between Two Rates. Journal of the American Statistical Association 50:1168-1194.

Retherford, R. D., and L. J. Cho. 1973. Comparative Analysis of Recent Fertility Trends in East Asia. Pp. 163-181 in International Union for the Scientific Study of Population (ed.), Proceedings of the 17th General Conference of the IUSSP, August 1973. Vol. 2. Liège, Belgium: International Union for the Scientific Study of Population.

Appendix Table 1.—Population Aged 14 Years and Older by Age (*I*), Region (*J*), Marital Status (*K*), and Sex (*L*): United States, 1940

J	*K*	*L*	*I* = 14-24	*I* = 25-34	*I* = 35-44	*I* = 45+
1[a]	1	1	6,077,395	1,864,473	804,881	1,077,309
2	1	1	2,231,505	524,485	226,033	399,335
3	1	1	3,090,353	625,359	253,080	419,171
1	2	1	767,715	4,048,445	4,328,747	7,373,783
2	2	1	370,162	1,597,307	1,472,711	2,455,969
3	2	1	396,285	1,351,121	1,387,888	3,108,490
1	3	1	85,707	324,938	460,291	1,705,112
2	3	1	37,167	113,235	143,841	645,567
3	3	1	34,372	71,611	87,322	592,583
1	1	2	5,538,511	1,499,021	706,200	1,117,055
2	1	2	1,731,657	297,398	134,495	247,419
3	1	2	2,118,950	253,371	110,181	181,608
1	2	2	1,785,075	4,676,236	4,250,469	5,731,361
2	2	2	860,972	1,753,808	1,367,892	1,895,980
3	2	2	860,715	1,503,101	1,441,182	2,390,921
1	3	2	207,030	599,014	846,383	3,815,524
2	3	2	70,527	146,017	198,655	1,133,611
3	3	2	62,990	90,086	112,969	812,792
Total			26,327,088	21,339,026	18,333,220	35,103,590
Grand total				101,102,924		

a- For an explanation of these category numbers, see the last section of the text.

Appendix Table 2.—Persons in the Labor Force by Age (*I*), Region (*J*), Marital Status (*K*), and Sex (*L*): United States, 1940

J	*K*	*L*	*I* = 14-24	*I* = 25-34	*I* = 35-44	*I* = 45+
1	1	1	3,077,342	1,707,329	706,852	718,406
2	1	1	1,088,644	438,263	164,911	204,606
3	1	1	1,778,602	572,784	226,616	312,910
1	2	1	744,377	3,973,494	4,215,065	6,138,377
2	2	1	358,946	1,563,217	1,424,570	1,911,770
3	2	1	385,299	1,328,310	1,358,939	2,786,518
1	3	1	69,357	287,721	403,894	965,244
2	3	1	26,821	82,134	100,745	292,919
3	3	1	28,813	62,079	76,688	361,678
1	1	2	2,400,191	1,293,244	575,797	579,371
2	1	2	512,254	214,531	84,412	86,707
3	1	2	407,490	116,490	39,662	39,325
1	2	2	385,603	1,030,946	770,353	568,925
2	2	2	98,418	258,181	192,616	162,336
3	2	2	57,777	103,939	87,641	90,331
1	3	2	122,277	422,525	559,159	897,820
2	3	2	31,422	81,443	103,098	206,504
3	3	2	21,017	39,941	52,682	150,831
Total			11,594,650	13,576,571	11,143,700	16,474,578
Grand total				52,789,499		

Appendix Table 3.—Population Aged 14 Years and Older by Age (I), Region (J), Marital Status (K), and Sex (L): United States, 1970

J	K	L	I = 14–24	I = 25–34	I = 35–44	I = 45+
1	1	1	11,708,759	1,509,681	679,413	1,371,986
2	1	1	3,315,162	321,785	165,719	422,472
3	1	1	712,049	73,149	42,940	127,322
1	2	1	2,479,694	6,846,531	6,860,497	15,988,504
2	2	1	775,569	2,275,328	2,223,362	5,107,775
3	2	1	57,846	242,136	388,784	1,347,368
1	3	1	441,681	702,882	714,339	3,026,023
2	3	1	110,210	178,194	173,014	894,550
3	3	1	11,164	18,117	20,630	147,269
1	1	2	10,218,862	1,052,330	545,479	1,859,618
2	1	2	2,628,069	178,925	103,411	352,084
3	1	2	557,317	30,188	17,916	60,862
1	2	2	4,024,955	7,211,102	6,937,038	13,929,463
2	2	2	1,305,839	2,449,820	2,203,546	4,364,818
3	2	2	103,614	306,312	444,886	1,164,627
1	3	2	817,755	1,201,362	1,328,989	9,269,213
2	3	2	205,512	228,904	260,583	2,220,179
3	3	2	24,773	19,201	22,275	266,458
Total			39,498,830	24,845,947	23,132,821	61,920,591
Grand total				149,398,189		

Appendix Table 4.—Persons in the Labor Force by Age (*I*), Region (*J*), Marital Status (*K*), and Sex (*L*): United States, 1970

J	*K*	*L*	*I* = 14-24	*I* = 25-34	*I* = 35-44	*I* = 45+
1	1	1	5,371,325	1,260,249	548,076	737,910
2	1	1	1,268,759	233,006	108,207	172,830
3	1	1	287,642	61,972	35,240	88,432
1	2	1	2,285,504	6,653,752	6,705,567	12,183,629
2	2	1	729,432	2,223,153	2,153,680	3,499,502
3	2	1	52,902	235,590	376,438	1,051,983
1	3	1	357,493	606,757	610,218	1,386,585
2	3	1	77,165	132,788	132,049	329,303
3	3	1	8,530	16,396	18,193	76,406
1	1	2	3,785,363	852,652	426,049	946,915
2	1	2	674,622	117,398	60,606	123,865
3	1	2	128,977	19,198	9,863	19,750
1	2	2	1,866,498	2,763,511	3,187,296	5,135,841
2	2	2	510,330	930,666	1,023,095	1,417,442
3	2	2	32,543	92,148	161,369	299,215
1	3	2	454,669	766,538	920,999	3,043,713
2	3	2	100,727	134,203	163,309	559,686
3	3	2	11,694	10,951	12,429	56,640
Total			18,004,175	17,110,928	16,652,683	31,129,647
Grand total				82,897,433		