

**ESTIMATION OF DEMOGRAPHIC PARAMETERS
FROM CENSUS AND VITAL REGISTRATION DATA**

Griffith FEENEY

East-West Population Institute
Honolulu, Hawaii

1. INTRODUCTION

In a paper presented to the 1969 conference of this association in London, Professor W. Brass observed that, ironically, one cause for dissatisfaction with techniques for the analysis of limited and defective population data has been the rapid improvement in the extent and accuracy of the data available. "The necessarily rigid techniques for imposing order on very restricted and doubtful information are too inflexible when materials are better. Further progress towards precise measurement can only be made by analyses which allow, to a greater degree, for the peculiarities of the particular populations ..." (1969 : 185). The present paper discusses strategy and tactics for attaining this analytical flexibility with particular reference to recent developments in the analysis of census data on fertility and child survivorship. The census data consists of tabulations of women's responses to questions on number of children born and number of these children surviving at the time of the census. The goal is to translate this information into estimates of age-specific fertility rates, or summary statistics thereof, and life table statistics of mortality. These fertility rates and life table statistics are directly calculable from vital registration data on number of births and deaths by age of mother and decedent, respectively, hence the present discussion is of interest in connection with populations for which vital registration is either nonexistent or seriously incomplete. Sections 2 to 7 give an overview of published developments of the past decade. Section 8 reports recent work of my own concerning the estimation of mortality trends from child survivorship data.

2. MODELS FOR THE AGE-SCHEDULE OF FERTILITY

A simple model for the age schedule of fertility is

$$m_x = \text{TFR} \times m'_x \quad (2.1)$$

where the m'_x are given constants with $\sum m'_x = 1$ and TFR is a parameter for the total fertility rate $\sum m_x$. This model is the basis for the calculation of so-called "indirectly standardized" birth rates and may also be used where available data provide some indication of the shape, but not the height, of the fertility schedule (see the following section for examples). Another model is defined by the "fertility polynomial" due to Brass

$$m(a) = \begin{cases} c(a - s)(s + 33 - a)^2 & \text{for } s \leq a \leq s + 33 \\ 0 & \text{otherwise} \end{cases} \quad (2.2)$$

(Brass 1975 : 18-23).

The total fertility rate corresponding to this schedule is $\text{TFR} = 98826.75c$ children per woman and the mean age at child-bearing is $\text{MAC} = s + 13.2$ years, hence the model is readily reparameterized in terms of the total fertility rate and the mean age at fertility. Numerical computations with the model are facilitated by the observation that $0.25(s + 33 - a)^4 - 11(s + 33 - a)^3$ is an antiderivative of the polynomial in (2.2) whence

$$\int_0^x m(a) da = c \times \left[\frac{(s + 33 - x)^4}{4} - 11(s + 33 - x)^3 + 98826.75 \right],$$

$$s \leq x \leq s + 33 \quad (2.3)$$

In particular, if women are uniformly distributed by age in the age group x to $x + n$,

$${}_n m_x = \int_x^{x+n} m(a) da = \int_0^{x+n} m(a) da - \int_0^x m(a) da, \quad (2.4)$$

which in concert with (2.3) expresses ${}_n m_x$ in terms of the parameters c and s . The model (2.4) therefore leads to the parameterization

$$m_x = \text{TFR} \times \phi_x(\text{MAC}), \quad (2.5)$$

of age-specific fertility rates m_x where $\phi_x = \frac{1}{m_x}$ as defined by (2.4) and (2.3), the latter with $c = 98826.75^{-1}$ and $s = MAC - 13.2$.

Observe that although (2.3) is valid only for x in the indicated range, substitution of $\min\{\max\{x,s\}, s + 33\}$ for x on the right side of the equals sign gives a formula valid for all x .

Coale and Trussell have recently proposed a more elaborate fertility model,

$$m(a) = TFR \times G(a; a_0, k) \times r(a; m) \quad (2.6)$$

where

$$G(a; a_0, k) = \frac{0.19465}{K} \int_0^x \exp \left\{ \frac{-0.174(a - a_0 - 6.6k)}{K} - \exp \left\{ \frac{-0.2881(a - a_0 - 6.06k)}{K} \right\} \right\} da \quad (2.7)$$

and

$$r(a; m) = Mn(a) \exp[mv(a)], \quad a \geq 0 \quad (2.8)$$

where $n(a)$ and $v(a)$ are empirically defined schedules and M is a constant chosen so that

$$\int_0^\infty m(a) da = TFR$$

(Coale and Trussell 1974). The schedule $n(a)$ represents marital fertility and the schedule $v(a)$ in conjunction with the parameter m represent the attenuation of fertility due to control of fertility within marriage. The function $G(a; a_0, k)$ is a model for the age distribution of women at first marriage (Coale and McNeil 1972). The parameters a_0 and k are essentially location and scale parameters, with a_0 representing the earliest age of marriage and k representing the extent to which marriages occur within a greater or lesser span of ages.

3. CENSUS DATA ON CUMULATIVE FERTILITY

Population censuses frequently query women as to the number of children they have borne in their lifetime. This information is generally tabulated by age of women, showing either the total number of children born to women in each age group or the distribution of women in each group by number of children born. The total number of children born may be calculated as the number of women who have had exactly one child, plus two times the number who have had exactly two children, and so forth, but it should be observed that this procedure is applicable only if the distribution is complete. In practice distributions are often truncated to conserve space in census publications. See Feeney 1976a for further discussion. Tabulations by quinquennial age groups are most common, but recent developments in computation and analysis are leading to wider production of single year tabulations.

Let N_x denote the number of women aged x in completed years at the time of a census and suppose all these women aged $x + 1/2$ exactly. The number of births to these women during the t -th year prior to the census may then be expressed as $N_x m_{x-t+1}(t)$, where $m_x(t)$ denotes the age-specific fertility rate for the one year age interval centered on exact age x for the t -th year prior to the census. Total births to these women equals the sum of these terms over all relevant values of t ,

$$B_x = \sum_{t=1}^{x-9} N_x m_{x-t+1}(t), \quad x = 10, \dots, 59 \quad (3.1)$$

where B_x denotes all children born to women aged x in completed years at the time of the census and the upper and lower limits of the fertile age span are set at exact ages 10 and 60, respectively.

The following formulas are easier to write down than to read. The trick is to imagine the terms of the sums in (3.1) written out as an array with rows corresponding to ages and columns corresponding to years prior to the census. All the formulas are derived by multiplying all terms in a row by a constant or by first combining rows by addition, which corresponds to age grouping, and multiplying the resulting rows by a constant.

Division of both sides of (3.1) by N_x gives

$$b_x = \sum_{t=1}^{x-9} m_{x-t+1}(t), \quad x = 10, \dots, 59, \quad (3.2)$$

where b_x denotes mean children born to women aged x in com-

pleted years. These equations are tautological if the age-specific fertility rates are understood to apply only to women surviving at the time of the census; otherwise they incorporate the assumption that women not surviving to the time of the census exhibit the same age-specific fertility rates as women who do survive.

Equations for age groups are obtained by reverting to (3.1), summing over successive single years of age, and dividing both sides of the result by the number of women at these ages.

$${}_n b_x = \frac{\sum_{y=x}^{x+n-1} \sum_{t=1}^{y-9} N_y m_{y-t+1}(t)}{\sum_{y=x}^{x+n-1} N_y}, \quad \begin{matrix} n=1, \dots, 60-x \\ x=10, \dots, 59 \end{matrix} \quad (3.3)$$

where ${}_n b_x$ denotes mean children born to women aged x to $x+n-1$ in completed years. Under the assumption that women are uniformly distributed in this age interval,

$$N_y = N, \quad y = x, x+1, \dots, x+n-1 \quad (3.4)$$

which upon substitution in (3.3) yields

$${}_n b_x = \frac{1}{n} \sum_{y=x}^{x+n-1} \sum_{t=1}^{y-9} m_{y-t+1}(t) \quad (3.5)$$

4. CONSTANT FERTILITY

Given census data for b_x , (3.2) defines a system of 50 equations in the $1 + 2 + \dots + 50 = 820$ unknowns $m_x(t)$. If fertility has been constant, however,

$$m_x(t) = m_x(1), \quad t = 1, \dots, 60 - x, \quad x = 10, \dots, 59 \quad (4.1)$$

and (3.2) becomes

$$b_x = \sum_{t=1}^{x-9} m_{x-t+1}, \quad x = 10, \dots, 59 \quad (4.2)$$

This defines a system of 50 equations in the 50 unknowns m_x , $x=10, 11, \dots, 59$, and has the simple solution

$$m_x = b_x - b_{x-1}, \quad x=10, \dots, 59 \quad (4.3)$$

This method has been developed by Mortara (1949 : 40-50).

Suppose next that age-specific birth rates calculated from vital registration data are available, and that, although births are underregistered the extent of underregistration is independent of age of mother. Alternatively, suppose that age-specific birth rates have been calculated on the basis of a census or survey question addressed to women inquiring whether or not they had given birth during the year or other period preceding the census and, if so, how many births; and assume that errors in the reported numbers of births, whether due to memory lapse, reference period error, or some other cause, are independent of age of women. In either case, the data specify the "shape" but not the "height" of the age-schedule of fertility, that is, the parameterization (2.1) applies, $m_x = \text{TFR} \times m'_x$, $x = 10, \dots, 59$. Entering these expressions for m_x in (4.2) gives

$$b_x = \text{TFR} \sum_{t=1}^{x-9} m'_{x-t+1}, \quad x=10, \dots, 59 \quad (4.4)$$

This system of 40 equations in the single unknown TFR will not in general have a solution, but the equations may be solved individually to give

$$\text{TFR}_x = b_x \div \left\{ \sum_{t=1}^{x-9} m'_{x-t+1} \right\}, \quad x=10, \dots, 59 \quad (4.5)$$

This dispersion and pattern of these solutions may be examined by plotting them against x , and a final estimate of the TFR determined as circumstances warrant. The idea of this method is due to Brass who has however developed it for application to data in five year groups as indicated below.

Under the constant fertility assumption (4.1) the double sum in (3.5) reduces to

$$n b_x = \sum_{i=10}^{x-1} m_i + \sum_{i=0}^{n-1} \frac{n-i}{n} m_x^i, \quad n=1, \dots, 60-x \quad (4.6)$$

$x=10, \dots, 59$

zero subscript is $x+i$

as may be seen simply by writing out the terms of the sum in array form. This formula applies for x and n beyond the indicated ranges taking $m_x=0$ for $x<10$ and $x\geq 60$. Substituting the parameterization $m_x = \text{TFR} \times m'_x$ for m_x ,

$${}_n b_x = \text{TFR} \times \left\{ \sum_{i=10}^{x-1} m'_i + \sum_{i=0}^{n-1} \frac{n-i}{n} m'_{x+i} \right\}, \quad n=1, \dots, 60-x \quad (4.7)$$

$$x=10, \dots, 59$$

which may be dealt with in the same manner as (4.4).

Equation (4.6) may be used where census data on children ever born are not available by single years of age, but it does require that the m_x be available by single years. Where this is not the case, group data may be interpolated to single years and the preceding formulas applied to the interpolated values. Alternatively, observe that the left summation in (4.7) may be calculated as

$$\sum_{(x,n) \in I} n^m x \quad (4.8)$$

where I is an index set defining age groups which cover all ages between exact age 10 and exact age x . The right summation in (4.7) is indeterminate but may be expressed as $k_n m_x$ where

$$k = n - n \left[\sum_{i=0}^{n-1} i \times \{ m_{x+i} \} \right] \quad (4.9)$$

If it be assumed that the m_x values here conform to the polynomial fertility model (2.5), the term in curly brackets depends only on the value of MAC and value of this "multiplier" may be tabulated for various age groups and values of MAC (Brass 1975:18-23). Values of MAC may be estimated by entering the parameterization (2.5) into the equations (4.7), which gives

$${}_n b_x = \text{TFR} \times {}_n \phi_x(\text{MAC}), \quad n=1, \dots, 60-x \quad (4.10)$$

$$x=10, \dots, 59$$

where

$${}_n \phi_x(\text{MAC}) = \sum_{i=0}^{x-1} \phi_x(\text{MAC}) + \sum_{i=0}^{n-1} \frac{n-i}{n} \phi_{x+i}(\text{MAC}), \quad (4.11)$$

$$\begin{cases} n=1, \dots, 60-x \\ n=10, \dots, 59 \end{cases}$$

The ratio of two such equations has the form

$$\frac{\frac{n^b}{b} x}{n^b x-n} = \frac{n^b \phi_x(\text{MAC})}{n^b \phi_{x-n}(\text{MAC})} \quad (4.12)$$

which may be readily solved by interpolation among tabulated values of the expression on the right, which depends only on MAC and not on the data. The MAC value obtained will of course depend to some extent on the age groups chosen and, as previously, it is useful in practice to examine the dispersion of the estimates obtained from various choices.

5. ESTIMATES FROM TWO CENSUSES

Suppose next that children born data are available for two successive censuses separated by a whole number of years. Let formula (3.2) refer to the later census and observe that the sum on the right splits into terms referring to fertility prior to the earlier census and terms referring to fertility between the two censuses.

$$b_x = \begin{cases} \sum_{t=1}^{x-9} m_{x-t+1}(t) & x = 10, \dots, T+9 \\ \sum_{t=1}^T m_{x-t+1}(t) + \sum_{t=T+1}^{x-9} m_{x-t+1}(t) & x = T+10, \dots, 59 \end{cases} \quad (5.5)$$

If no women who would be of reproductive age at the later census die during the intercensal period, the term at bottom right represents mean children born to women aged $x-t$ in completed years at the earlier census. This leads to

$$b_x = \sum_{t=1}^T m_{x-t+1}(t) + b'_{x-T}, \quad x=10, \dots, 59+T \quad (5.6)$$

where b'_x denotes mean children born values for the earlier census. Taking $b'_x=0$ for $x<10$ and $m_x(t)=0$ for $x \geq 60$ for all t , (5.6) defines a system of 60 equations in the $50 \times T$ unknowns $m_x(t)$. This equation is not in general directly solvable unless $T=1$, the uninteresting, because practically nonexistent, case of two censuses one year apart. Two options

are open, however. Intercensal b_x values at single year intervals might be interpolated from the pairs of census values, and the various single year systems so generated solved directly, a procedure which has been proposed by Ansley Coale following a similar strategem developed by Norman Ryder in connection with nuptiality (Coale : personal communication). Alternatively, one may observe that the sums in (5.6) are fertility rates for T-year age groups and that the totality of these rates generated as x ranges from 10 to 59+T divides into T groups each of which covers the fertile age span completely and without overlap. While the individual rates cannot be readily compared, a total fertility rate may be calculated from each subset. This procedure is obvious enough, and if it has not been proposed in the literature the reason is in all likelihood the relative rarity of census tabulations of children ever born by age. The advantage, and it is a real advantage, of single year tabulations is that mean children born to woman aged x in completed years at the time of the census closely approximates mean parity for exact age $x+1/2$, the midpoint of the age interval. The disadvantage of the conventional five year age groups, and it is a real disadvantage, is that this midpoint approximation can be seriously in error. Arretx has devised an ingenious procedure for adjusting quinquennial values to obtain estimates of mean parity at the midpoint of the intervals (1973).

6. UTILIZATION OF BIRTH REGISTRATION DATA

Suppose for the moment that available birth registration data, though suffering from underregistration of births, provides numerators for single year age-specific fertility rates for a substantial number of years prior to the census. Corresponding denominators may be obtained by reverse survival of the census age distribution. This requires estimates of adult mortality, but these estimates can be obtained in various ways, and where mortality is low, the reverse-survived values will be relatively insensitive to errors in the estimates. Formula (3.2) may be rewritten as

$$b_x = \sum_{t=1}^{x-9} m_{x-t+1}(t) \psi_x(t), \quad x=10, \dots, 59 \quad (6.1)$$

where $m_x(t)$ denotes the age-specific birth rate for age x in completed years for the t-th year prior to the census calculated from registered births and $\psi_x(t)$ denotes the ratio of total to registered births for this age and time period. Assume now that underregistration, as measured by $\psi_x(t)$ was

independent of age of mother and declined linearly during the years preceding the census, that is,

$$\psi_x(t) = u + r(t-1), t=1, 2, \dots, \quad (6.2)$$

where u denotes the level of underregistration during the year preceding the census and r denotes the rate of decline of this level in preceding years. Entering these expressions for $\psi_x(t)$ in (6.1) gives

$$b_x = \sum_{t=1}^{x-9} m_{x+t-1}(t) [u+r(t-1)], x=10, \dots, 59 \quad (6.3)$$

or, rearranging terms,

$$b_x = u \left[\sum_{t=1}^{x-9} m_{x+t-1}(t) \right] + r \left[\sum_{t=1}^{x-9} (t-1) m_{x+t-1}(t) \right], x=10, \dots, 59 \quad (6.4)$$

This constitutes a system of 40 equations in the 2 unknowns u and r and any two of these may be solved for u and r . This method has been developed and applied to Yugoslavia by Macura (1972:9-10 and throughout).

7. CHILD SURVIVORSHIP METHODS FOR CONSTANT MORTALITY

Censuses frequently query women on the number of children they have borne who are surviving at the time of the census as well as on the total number born, allowing the calculation of the proportion of deceased children among all children born to women in each age group. Disaggregating total children born into birth cohorts and surviving each cohort forward to the time of the census levels to the equation

$$Q_i = 1 - \sum_{t=1}^{n(i)} p(t)c_i(t), i=1, 2, \dots, N \quad (7.1)$$

where the index i specifies age group, Q_i denotes the proportion of all children born to women in this age group who are deceased at the time of the census, $c_i(t)$ the proportion of all children born who were born during the t -th year prior to the census and $p(t)$ the proportion of these children who survive to the time of the census. The propor-

tions $c_i(t)$ may be calculated directly from the terms of the sum in (3.1) if the age-specific fertility rates $m_x(t)$ are known, and these rates may be estimated by assuming^x constant fertility and fitting a model fertility schedule to the available data on mean children born to women in each age group, as for example by (4.12). The values $p(t)$ may be expressed as

$$\begin{aligned} p(1) &= L_0(1) & (7.2) \\ p(2) &= L_0(2) \times \frac{L_1(1)}{L_0(1)} \\ p(3) &= L_0(3) \times \frac{L_1(2)}{L_0(2)} \times \frac{L_2(1)}{L_1(1)} \end{aligned}$$

where $L_x(t)$ denotes person years lived between exact age x and exact age $x+1$ in the life table expressing mortality during the t -th year prior to the census. Inserting these expressions in (7.1) yields one equation for each age group in a total of $n(n+1) \div 2$ unknowns $L_x(t)$, $t=1, 2, \dots, N-x$, $x=0, 1, \dots, N-1$.

If mortality is assumed constant,

$$\begin{aligned} L_x(t) &= L_x, \quad t=1, \dots, N-x \\ &x=0, 1, \dots, N-1 \end{aligned} \quad (7.3)$$

the terms on the left in (7.2) cancel leaving $p(t) = L_{t-1}$, $t=1, 2, \dots, N$, which reduces the number of unknowns to N . This is still far from sufficient for solution, however, since N will be the upper limit of the oldest age group less the age at first childbearing, 35, for example, if the age groups are 15-19, ..., 45-49 and childbearing begins at 15 years of age. Introducing the further assumption that the age schedule of mortality conforms to a one-parameter model life table family, reduces (7.1) to

$$Q_i = 1 - \sum_{t=1}^{n(i)} L_{t-1}(\omega) c_i(t), \quad i=1, 2, \dots, N \quad (7.4)$$

where $L_x(w)$ denotes person years lived between exact age x and exact age $x+1$ in the model life table identified by the parameter value w .

Each of these equations may be solved independently for ω , yielding values $\omega_1, \dots, \omega_n$, say. Any desired life table

statistic may be calculated from these ω values using the defining formulas of the model life table family.

Brass developed an ingenious procedure for deriving mortality estimates which circumvents solving (7.4) directly. By the constant mortality assumption, all the children born have experienced mortality represented by a single life table, and by the model life table family assumption, this life table corresponds to the model life table with parameter ω for some ω . It follows that $q(x) = q(x;\omega)$, where $q(x) = 1 - l_x$ denotes the probability of death between birth and age x in the life table experienced by the population and $q(x;\omega)$ denotes the corresponding value in the model life table identified by the parameter value ω . It follows trivially from (7.4) that $q(x) = Q \times M(x,i,\omega)$ where the "multiplier" M is set equal to

$$\frac{q(x;\omega)}{n(i)} \quad (7.5)$$

$$1 - \sum_{t=1}^{x-i} L_{t-1}(\omega) c_i(t)$$

This quantity depends on the age group, represented by the index i , on the values of $c_i(t)$, and on the value of x , all of which are known or estimated. It also depends on ω however, that is, on the level of mortality, hence introduction of the multiplier does not obviously advance the cause of mortality estimation. But Brass discovered that if x is suitably chosen in relation to i , M is approximately constant with respect to changes in ω , hence that the value of the multiplier may be calculated approximately without knowledge of the level of mortality. He proceeded to table these multipliers for values of $c_i(t)$ obtained by fitting the model (2.2) to observed fertility, assumed constant (Brass and others 1968:105-120).

Brass utilized the ratio of mean children born for women 15-19 to mean children born for women 20-24 to fit model (2.2) using (4.12), hence the multipliers produced were functions of this ratio, usually referred to as a "mean parity ratio". Sullivan investigated this relationship by calculating exact values of both the multiplier (7.4) and the mean parity ratio, using (3.5), for all possible combinations of 65 observed fertility schedules and 40 model life tables, generating a total of $65 \times 40 = 2600$ values for both quantities. He then calculated regressions of the multipliers on the mean parity ratios for various subsets of these observations (Sullivan 1972:82-83). He also developed similar regression results for use with proportions of deceased children among all children born to women classified by duration of marriage, results which have seen relatively little use for want of suitable data (see however Sishawy 1975). Sullivan's regression results

for age improve on the original Brass procedure by substituting observed fertility schedules for the Brass polynomial model (2.2). Trussell subsequently refined the procedure still further by including further independent variables in the regression and substituting model schedules derived from the Coale-Trussell model (2.6-8) for Sullivan's observed schedules (Trussell 1975). The significance of the latter point is that accurate single-year age-schedules of fertility are with rare exceptions available only for populations in which childbearing begins late, hence that Sullivan's results are of questionable applicability in populations in which childbearing begins early. The Coale-Trussell model generates an early age at childbearing model age-schedule of fertility and so partially fills the gap in available data. Observed schedules would be preferable, to be sure, but the excellent fit of the Coale-Trussell model to diverse empirical schedules suggests the derived schedules for early age at childbearing are in fact good approximations to actual schedules.

8. MORTALITY TRENDS FROM CHILD SURVIVORSHIP

Suppose next that mortality has been changing during the years preceding the census but that the life table representing mortality at any time prior to the census conforms to a one-parameter model life table family, and suppose further that the trend of this parameter over time is linear. Although the parameter may be considered to vary continuously with time, it will suffice to consider the life table applicable at the midpoint of each year as applicable throughout the year. Then the parameter value for the t -year prior to the census is $\omega_t = \omega + (t-1/2)r$, $t=1, 2, \dots, k$, and we set $L_x(t)$ in (7.2) equal to L_x in the model life table identified by the parameter value ω_t . This defines each $p(t)$ in the basic equation (7.1) as a function of ω and r , whence we arrive at

$$Q_i = 1 - \sum_{t=1}^{n(i)} p(t; \omega, r) c_i(t), \quad i=1, \dots, N \quad (8.1)$$

The expression $p(t; \omega, r)$ denotes here the proportion of persons born during the t -th year prior to the census who would survive to the time of the census given that the model life table parameter was changing at a constant annual rate r during the years preceding the census and had the value ω at the time of the census.

One might attempt to find values for ω and r for which the values on the left in (8.1) are "close to" the

observed proportions Q_i or to solve subsystems of pairs of equations for ω and r . Another possibility is to consider the "solution set" of each equation in (8.1), that is, the set of all combinations of values for r and ω which satisfy the equation. The points of the solution set may be represented as $(r, \Omega(r))$, where r ranges over all permissible values of the rate of change and $\Omega(r)$ denotes the solution for ω of the equation formed by holding r constant at the indicated value. The solution set may be visualized as the graph of the function Ω . Given a finite series of r -values r_1, \dots, r_m one may calculate corresponding ω -values $\omega_1, \dots, \omega_m$, and plotting the points $(r_1, \omega_1), \dots, (r_m, \omega_m)$ gives a finite approximation to this graph. Intermediate values may be obtained by interpolation.

The solution set of the i -th equation in (8.1) defines a family of linear trends in the model life table parameter consistent with observed child survivorship in the i -th age group. Observed child survivorship Q_i for a given group might lead us to conclude, for example, that if mortality had been constant prior to the census, then the infant mortality rate must have been 197.2 infant deaths per thousand births; that if the infant mortality rate had been declining at the rate of $r=2$ infant deaths per thousand births per year, then the infant mortality rate at the time of the census must have been 195.0; that if the rate of decline had been $r=4$ infant deaths per thousand births per year, then the infant mortality rate at the time of the census must have been 192.7; and so forth for any specified series of r -values. These values were in fact obtained from child survivorship data from the 1960 census of North Borneo, for which, for the 15-19 age group, $Q = 0.189$ and the values of $c(t)$ were estimated to be 0.501, 0.302, 0.148, 0.046 and 0.003 for $t = 1, \dots, 5$, and zero for $t > 5$. The data are shown in Table 1. The model life table family used to define the function $p(t; \omega, r)$, $t = 1, \dots, 5$, via (7.2) was defined by the Brass logit transformation, reparameterized by the infant mortality rate and with β set constant at unity, applied to the Brass general standard l_x values (Brass 1971:69-110). Specifically,

$$L_x(\omega) = \begin{cases} 0.3 + 0.7l_1(\omega) & \text{if } x=0 \\ 0.5 [1_x(\omega) + l_{x+1}(\omega)] & \text{if } x=1, 2, \dots \end{cases} \quad (8.2)$$

where

$$l_x(\omega) = \left[1 + \exp \{ [A(\omega) + 0.51n(1-l_x) \div l_x] \} \right]^{-1} \quad (8.3)$$

observed proportions Q_i or to solve subsystems of pairs of equations for ω and r^i . Another possibility is to consider the "solution set" of each equation in (8.1), that is, the set of all combinations of values for r and ω which satisfy the equation. The points of the solution set may be represented as $(r, \Omega(r))$, where r ranges over all permissible values of the rate of change and $\Omega(r)$ denotes the solution for ω of the equation formed by holding r constant at the indicated value. The solution set may be visualized as the graph of the function Ω . Given a finite series of r -values r_1, \dots, r_m one may calculate corresponding ω -values $\omega_1, \dots, \omega_m$, and plotting the points $(r_1, \omega_1), \dots, (r_m, \omega_m)$ gives a finite approximation to this graph. Intermediate values may be obtained by interpolation.

The solution set of the i -th equation in (8.1) defines a family of linear trends in the model life table parameter consistent with observed child survivorship in the i -th age group. Observed child survivorship Q_i for a given group might lead us to conclude, for example, that if mortality had been constant prior to the census, then the infant mortality rate must have been 197.2 infant deaths per thousand births; that if the infant mortality rate had been declining at the rate of $r=2$ infant deaths per thousand births per year, then the infant mortality rate at the time of the census must have been 195.0; that if the rate of decline had been $r=4$ infant deaths per thousand births per year, then the infant mortality rate at the time of the census must have been 192.7; and so forth for any specified series of r -values. These values were in fact obtained from child survivorship data from the 1960 census of North Borneo, for which, for the 15-19 age group, $Q = 0.189$ and the values of $c(t)$ were estimated to be 0.501, 0.302, 0.148, 0.046 and 0.003 for $t = 1, \dots, 5$, and zero for $t > 5$. The data are shown in Table 1. The model life table family used to define the function $p(t; \omega, r)$, $t = 1, \dots, 5$, via (7.2) was defined by the Brass logit transformation, reparameterized by the infant mortality rate and with β set constant at unity, applied to the Brass general standard l_x values (Brass 1971:69-110). Specifically,

$$L_x(\omega) = \begin{cases} 0.3 + 0.7l_1(\omega) & \text{if } x=0 \\ 0.5 [1_x(\omega) + l_{x+1}(\omega)] & \text{if } x=1, 2, \dots \end{cases} \quad (8.2)$$

where

$$l_x(\omega) = \left[1 + \exp \{ [A(\omega) + 0.51n(1-l_x) \div l_x] \} \right]^{-1} \quad (8.3)$$

and

$$A(\omega) = 0.51n \{ [1 \div (1-\omega)] - 1 \} - 0.51n(1-\ell_1) \div \ell_1 \quad (8.4)$$

with ℓ_x values 0.8499, 0.8070, 0.7876, 0.7762, 0.7691 for $x=1, \dots, 5$, respectively.

TABLE 1
CHILD SURVIVORSHIP DATA FOR NORTH BORNEO :
CENSUS OF AUGUST 11, 1960

Age Group	Total Women	Children Born	Children Surviving
15-19	19,945	4,349	3,527
20-24	18,175	26,121	21,285
25-29	19,507	59,454	46,733
30-34	14,273	59,733	45,929
35-39	13,809	68,212	50,083
40-44	10,404	54,479	38,505
45-49	8,071	42,840	29,282
50-54	6,207	31,673	20,571
55-59	3,483	17,898	11,395
60-64	3,624	17,750	10,780
65-69	1,798	9,024	5,393
70-74	3,131	15,171	8,041

Source : North Borneo, *Report on the Census of Population Taken on 10th August, 1960* (Kuching, Sarawaki:Government Printing Office March 1962). Table 5, page 140 for total women and table 16, page 234 for children born and children surviving.

The solution set approach does extract information on the trend of mortality from the child survivorship data by ruling out some trends as inconsistent with the data, but the information takes an evidently awkward and impractical form. I proceeded to make the calculations described in the preceding paragraph nonetheless, simply out of curiosity, and without any particular expectation graphed the several consistent trends together. The result, shown in Figure 1, came as a profound if not ultimately unpleasant shock : the consistent trends intersect precisely at a single point a certain number

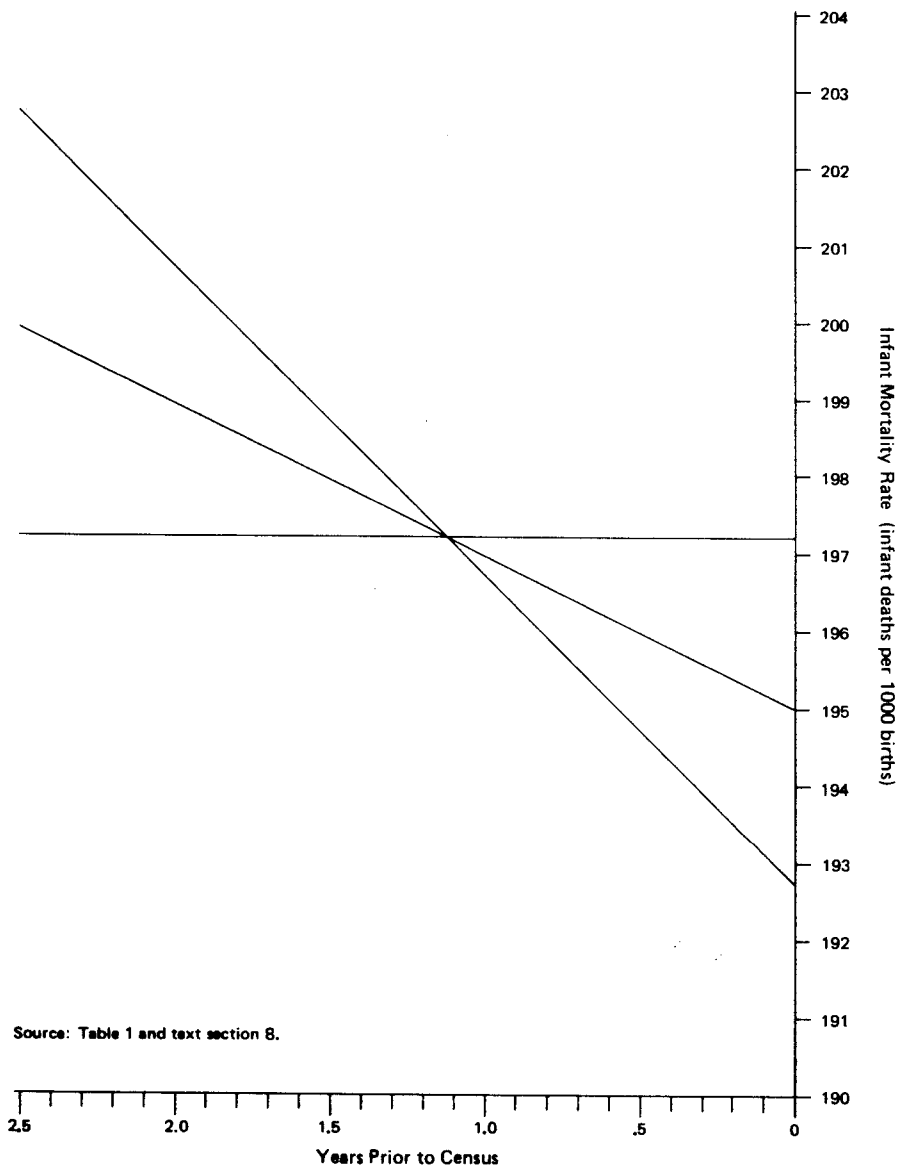


FIGURE 1

Linear infant mortality trends consistent with child survivorship among children born to North Borneo women aged 15-19 as of the 1960 Census

of years prior to the census. A moment's reflection may be necessary to appreciate the significance of this fact. It means that calculation of the family of consistent trends provides a precise estimate of the level of mortality a certain number (not necessarily integral) of years prior to the census, even though one never learns which of the consistent trends was actually experienced by the population. Moreover, since one consistent trend is always constant mortality, the estimated infant mortality rate this number of years prior to the census must equal the rate estimated on the assumption that mortality has been constant. Finally, since under the one-parameter model life table assumption the infant mortality rate determines all values of the life table, the trends of all other life table statistics consistent with observed child survivorship, though not necessarily linear, must intersect at the same number of years prior to the census as the infant mortality trends, hence the solution set provides an estimate of any desired life table parameter this number of years prior to the census. These propositions are of course valid under the stipulated assumptions concerning the mortality trend.

I have subsequently calculated consistent linear trends for all age groups available in twenty different censuses representing Cambodia, Fiji, Indonesia, Malaysia, Papua New Guinea, Philippines, Thailand, American Samoa, Brunei, North Borneo and Sabah (now part of Malaysia), Korea, and Gilbert and Ellice Islands Colony. The common intersection of consistent linear trends occurs generally and to a numerical precision which is completely negligible (of the order of one half of one percent) for the younger age groups and rises to at most 5 percent for women aged 70-74 at the 1960 census of North Borneo. Whether the intersections are only close approximations or are exact with apparent discrepancies due to errors of computation I do not know, and intriguing a theoretical question as this is, lack of an answer in no way impairs the use of the results for mortality estimation. The years-prior-to-census values increase with age group of women, by about 2 1/2 years with each successive quinquennial age group. The values for the North Borneo data in Table 1, for example, are 1.12, 2.64, 5.28, 7.50, 10.31, 13.55, 17.00, 20.27, 27.07, 25.61, and 27.49 years for age groups 15-19, ..., 70-74, respectively. Consequently, although the consistent trends for any single age group provide only a level of mortality at a single time, with no indication of which trend was experienced, the data for the age groups taken together provide a series of estimates over a period of 2-3 decades prior to the census. This evidently provides an internal consistency check on the linearity assumption, incidentally, since the calculations for each age group are entirely independent. Examination of the graphs of the estimated infant mortality rates over time suggests strongly that the results cannot all or always be

accepted at face value. The estimate based on the 15-19 age group indicates a rise in infant mortality prior to the census in virtually every case, including cases of two and three successive censuses of the same population. There can be little doubt that this rise is spurious. A less pronounced but still clearly visible pattern is an apparent rise in mortality during the third and sometimes the second decade preceding the census, probably in most cases, a spurious indication resulting from deteriorating data for advanced age groups. The results nonetheless appear to provide a useful picture of mortality for two and in some cases nearly three decades prior to the census.

9. CONCLUDING REMARKS

Three principal stages are discernable in the estimation techniques discussed here, the formulation of tautological equations relating the desired unknown quantities and the available data, the introduction of parameterization, resulting in derived equations with fewer unknowns than the original tautological equations, and the final solution of these derived equations. The original tautological equations, such as (3.2), (6.1) or (7.1), are severely underdetermined, that is, many combinations of values for the desired unknown quantities are equally consistent with the available data. This is to be expected and might indeed be said to be the defining characteristic of insufficient data problems. If the original tautological equations have a unique solution for the desired unknown parameter values in terms of the available data, the data can hardly be said to be "insufficient", though one might be faced with an "inconvenient data problem" if the solution is difficult to obtain.

The solution set of the original tautological equations may be said to exhaust the information contained in the data relevant to the determination of the desired unknown quantities. Specific values for these quantities (or a "smaller" solution set) can be obtained only by introducing further information into the problem. Since available data may be supposed to have been exhausted in the formulation of the tautological equations, this amounts to asking whether, on the basis of circumstantial evidence or empirical regularities observed in other populations, some points of the solution set are more likely than others as candidates for the desired unknown quantities. This information is introduced into the problem by a series of parameterization equations, which express several of the unknown quantities in the problem as functions of smaller numbers of formal parameters. The families of model fertility schedules defined by (2.1), (2.2)

and (2.3-5) and the model life table families used in sections (7) and (8) are examples of parameterization based on empirical regularities, whereas (4.1), (6.2) and (8.1) represent assumptions about time trends of various quantities which might be justified by circumstantial evidence.

Substitution of the parametric expressions in the parametric equations for the corresponding unknowns in the original tautological equations results in a derived system of equations with a smaller number of unknowns. Examples of derived equations occur at (4.2), (4.7), (6.4), (7.4) and (8.1). Since parameterization reduces the number of unknowns in the problem, the derived equations may have a unique solution or may be overdetermined, but they will in any case be less underdetermined than the original tautological equations. From the mathematical point of view, the sole purpose of parameterization is to reduce the number of unknowns and so obtain a more tractable set of equations. From the demographic point of view parameterization represents information introduced into the problem beyond that contained in the available data. In the nature of the situation, it is impossible to directly verify the information so introduced, and parameterization must therefore be considered as a source of error in the final estimates along with errors due to errors in the original data.

The final stage consists of solving the derived equations for the unknown parameter values and calculating the corresponding values of the originally desired unknown quantities from the parametric equations. Thus, for example, one solves (7.4) for ω and then calculates, say $q(5)$ from ω using the formulas defining the model life table family. The latter step may in some cases be degenerate, however, as at (4.2-3). Where the derived equations are over-determined, one looks for "best-fit" solutions.

These steps may be observed, with various individual idiosyncrasies, in virtually every technique associated with estimation from "limited" or "insufficient" data.

With respect to the development of techniques in the future, it is evident that the possibilities of "weak" parameterization, that is, parameterizations involving greater numbers of unknown parameters (the mathematical point of view) and less restrictive demographic assumptions (the demographic point of view) have hardly begun to be explored. Methods currently in use typically reduce problems to finding a single unknown parameter. Virtually no use is made of evaluations-only methods, as opposed to derivative methods, for simultaneous systems of nonlinear equations and numerical minimization for "best fit" solutions, and while it is true that difficulties tend to increase with the square or the cube of the number of unknowns, modern computations facilities render two- and three-dimensional problems well within reach.

With respect to tabulation, one point is too obvious to escape mention, and this is the overwhelming desirability of producing tabulation by single years. The elaborate analytical circumlocution necessitated by grouped data in connection with fertility estimation, (4.8-12), may be unavoidable where exceedingly small numbers are involved, but this will never be the case with census and vital statistics data, and often not with large-scale survey data. Grouping is, to be sure, a useful analytical device, but only if one has the option of which groups to introduce, and this means basic tabulation by single years.

RÉSUMÉ

Estimation de paramètres démographiques à partir de données du recensement et de l'état civil

Cette communication fait un tour rapide des méthodes d'estimation des taux de mortalité et de fécondité par âges, qui utilisent les données du recensement ou d'enquête fournissant le nombre d'enfants nés par femme et le nombre d'enfants survivants à la date de l'enquête ou du recensement. Il faut souligner que l'apparente diversité de ces méthodes renferme une similitude remarquable des approches, à un point tel que chacune de ces méthodes peut être considérée comme la réalisation, dans un contexte particulier, d'une approche stratégique unique aux problèmes d'estimation. On pense que la reconnaissance explicite de cette approche pourrait s'avérer utile pour des développements ultérieurs des méthodes existantes et pour l'élaboration de nouvelles méthodes.

REFERENCES

- Arretx, C., 1973, "Fertility estimates derived from information on children ever-born using data from successive censuses", *International Population Conference : Liège 1973*, Vol. 2, pp. 247-261, Liège : IUSSP.
- Brass, W., 1971a, "Disciplining demographic data", *International Population Conference : London 1969*, Vol. 1, pp. 183-204, Liège : IUSSP.
- 1971b, "On the scale of mortality", W. Brass, Ed., *Biological Aspects of Demography*, New York : Barner and Noble.
- 1975, "Methods of estimating fertility and mortality from limited and defective data". Based on seminars held 16-24 September 1971 at the Centro Latino-americano de Demographia (CELADE), San Jose, Costa Rica. An Occasional Publication of the Laboratories for Population Statistics, University of North Carolina at Chapel Hill, Chapel Hill, N.C.
- Brass, W. and others, 1968, *The Demography of Tropical Africa*, Princeton, New Jersey : Princeton University Press.
- Coale, A.J. and Trussell, T.J., 1974, "Model fertility schedules : variations quency of first marriage in a female cohort", *Journal of the American Statistical Association* 67 (340, December), 743-749.
- Coale, J. and Trussell, T., 1974, "Model fertility schedules : variations in the age structure of childbearing in human populations", *Population Index* 40(2), 185-258.
- Feeney, G., 1976a, "Tabulation of census and survey data on child survivorship", *Asian and Pacific Census Newsletter* 3(1), 5-6. Available from the East-West Population Institute, Honolulu, Hawaii.
- 1976b, "Estimating infant mortality rates from child survivorship data by age of mother", *Asian and Pacific Census Newsletter* 3(2), 12-16.
- Mortara, G., 1949, "Methods of using census statistics for the calculation of life tables and other demographic measures" (with applications to the population of Brazil) *Population Studies*, No. 7, Department of Social Affairs, United Nations, New York.

- Macura, M., 1972, "Estimates of the completeness of registration of births and infant deaths in Yugoslavia and its main provinces from the late 1940's to 1961", Ph.D. Dissertation, Department of Economics, Princeton University (available from Xerox University Microfilms, Ann Arbor, Michigan).
- McNeil, D.R. and Tukey, J.W., 1975, "Higher-order diagnosis of two-way tables, illustrated on two sets of demographic empirical distributions", *Biometrics* 31 (June), 487-510.
- Sishawy, N.N.el., 1975, "Measures of fertility and mortality in governance of Egypt 1947 and 1960", Ph.D. Dissertation, Office of Population Research, Princeton University.
- Sullivan, J., 1972, "Models for the estimation of the probability of dying between birth and exact ages of childhood", *Population Studies* 26(1), 116-133.
- Trussell, T.J., 1975, "A re-estimation of the multiplying factor for determining childhood survival", *Population Studies* 29(1), 97-107.
- Tukey, J., 1962, "The future of data analysis", *Annals of Mathematical Statistics*, 33, 1-67.

- Macura, M., 1972, "Estimates of the completeness of registration of births and infant deaths in Yugoslavia and its main provinces from the late 1940's to 1961", Ph.D. Dissertation, Department of Economics, Princeton University (available from Xerox University Microfilms, Ann Arbor, Michigan).
- McNeil, D.R. and Tukey, J.W., 1975, "Higher-order diagnosis of two-way tables, illustrated on two sets of demographic empirical distributions", *Biometrics* 31 (June), 487-510.
- Sishawy, N.N.el., 1975, "Measures of fertility and mortality in governance of Egypt 1947 and 1960", Ph.D. Dissertation, Office of Population Research, Princeton University.
- Sullivan, J., 1972, "Models for the estimation of the probability of dying between birth and exact ages of childhood", *Population Studies* 26(1), 116-133.
- Trussell, T.J., 1975, "A re-estimation of the multiplying factor for determining childhood survival", *Population Studies* 29(1), 97-107.
- Tukey, J., 1962, "The future of data analysis", *Annals of Mathematical Statistics*, 33, 1-67.