

## **DATA QUALITY**

Experience with the application of child survivorship estimation techniques shows that the quality of child survivorship data is extremely variable, ranging from very good to extremely poor.<sup>4</sup> [<sup>4</sup>Data for many countries and time periods are conveniently collected in United Nations (1979)] Data quality varies between national populations, within national populations over time, and between subgroups of national populations at a given time. Thus, Western Samoa data for the censuses of 1956, 1961, and 1966 appear to be reasonably good, whereas the data for the 1911 and 1976 censuses are, by comparison with earlier data and on the grounds of plausibility, so bad as to be useless for mortality estimation (Banister 1979). Peninsular Malaysia data for 1970 are, by comparison with vital registration data, quite good for Malay and Indians and extremely bad for Chinese.<sup>5</sup> [<sup>5</sup>The 1970 Peninsular Malaysia data on children surviving have never been published, but have been made available to researchers.]

Cases of really bad data reveal themselves by their own absurdity, but there is only limited comfort in this, for if some data can be so bad as to be obviously useless, it is only reasonable to suppose that other data will be sufficiently bad to mislead and yet not bad enough to betray itself.

Having said the worst, it should be emphasized that the situation is not nearly so bad as the worst suggests it could be. Reasonably good data seem to be more common than very bad data, and in most situations there will be sufficient information available to test data quality and so guard against misleading conclusions.

## **ESTIMATION VERSUS DIAGNOSIS**

There is nothing remarkable in calling attention to the need to consider errors in demographic data. Indeed, this is a truism. But it is frequently suggested, more by implication and example than by direct statement, that the calculation of estimates should not be undertaken until the quality of the data on which they are based has been assessed. The correct sequence is in fact precisely the opposite. The first step is to calculate estimates. The second step is to analyze the estimates, comparing them with other sources of information. The calculation of estimates comes first because the transformation of the data into estimates is by far the best way we have to assess data quality.

The practical conclusion is that the “estimation” techniques are tools for diagnosis as well as for estimation. We calculate estimates first to diagnose data quality, and then proceed to conclusions about levels, trends, and differentials in mortality, using the results of the diagnosis stage to establish margins of error.

## **LESSONS OF EXPERIENCE**

The mechanical application of estimation methods to data is ultimately a fairly trivial matter, though considerable effort is required to master the relevant literature. What matters most in practice is the ability to use the methods to analyze data. How are the techniques used to analyze data? How do we diagnose data quality? How do we evaluate

the validity of the assumptions? What is the net effect on the estimates? How much can we conclude, and how reliably, about demographic reality from the available data? These are matters of judgment in which experience as well as Knowledge of the particular method used plays a key role, We have a good deal of experience with applications for scores of populations, and a number of lessons stand out.

#### *Calculate Estimates from Data for All Age Groups*

It is widely believed that reports of numbers of children born and surviving are unreliable for women past middle reproductive ages. Thus, Sullivan (1972) provide, multipliers only for age groups 15—19, 20—24, and 25—29, forcing the user to ignore data for women over age 30. The reason given for this belief is that mean values of children ever born decline with advancing age beyond a certain age. This is sometimes considered to reflect memory lapse among older women. If reporting of children born and children surviving decline, however, the effect. will partially cancel out in the calculation of proportions of deceased children. In any event, numerous application, show that data for women up to and even over age 50 give reasonably good estimate. (Feeney 1977a).

At the same time, other applications show that data for young women may be very bad, a phenomenon that presumably cannot be explained by memory lapse (Cho and Feeney 1976). In view of these facts, the sensible rule is to abandon preconceptions and let the data speak for itself. Calculate estimates for all available age groups, scrutinize the results carefully, making the comparisons indicated below, and only then decide what should be believed and what should be discounted or discarded.

#### *Compare Estimates for Subpopulations*

Peninsular Malaysia provides a good example of the value of this rule. The estimates obtained for the total population from the 1970 census data are substantially low by comparison with vital registration data that is reasonably complete. When estimates are calculated for community groups, however, the estimate. for Malay. and Indians are found to be quite good, and the estimates for Chinese extremely bad. The errors in the total population estimates are thus largely because of the bad data for the Chinese subpopulation. Incidentally, this indicates that the data problems in this instance lie with the respondents rather than with the census operation.

#### *Compare Estimates from Successive Censuses*

American Samoa, Fiji, Gilbert and Ellice Islands, and Hungary provide examples in which estimates from two successive censuses provide evidence in favor of both data quality and the validity of the assumption. used (Feeney 1976b and 1977.). The East Malaysian state of Sabah provides an example of how such comparisons can generate a warning that something is seriously wrong (Feeney 1977a). Korea provides an example showing that consistency does not prove correctness, for three successive censuses yield reasonably consistent estimates, all of which are seriously low (Cho and Feeney 1976).

### *Compare Estimates with Vital Registration Data*

The Philippines provides an example showing that this comparison can be very useful even if the vital registration data are substantially incomplete (Feeney 1977a). The infant mortality rates estimated from Philippines child survivorship data are substantially lower than rates calculated from vital statistics, which are almost certainly below the true value.. The conclusion is that the child survivorship data is seriously flawed and probably useless for mortality estimation.

### *Compare Estimates with Birth History Data*

There seems to be little doubt that, short of a fully developed vital registration system, birth historic, fro. a well—conducted fertility survey provide the best estimates of infant and child mortality, at a substantial cost, of course, and for relatively little data, as the numbers will be too small to allow much disaggregation by areas or characteristics. Birth history data probably provides the best information on the appropriate model life table specification. An analysis of data for Nepal illustrating this lesson is given in Thapa and Retherford (1982).

### *Discount Estimates from the 15—19 Age Group*

Children born to women aged 15—19 were obviously born to mothers less than age 20 at the time of birth, and births to young women experience relatively high mortality. This biases the mortality estimates from this age group upward. The bias shows clearly in one application after another and is confirmed by direct data on infant mortality rates by age of mother (Feeney 1980). Data for women aged 20-24 are slightly biased as well. Ewbank (1982) has recently suggested a method of adjusting for this bias and has applied it to data for Bangladesh. It appears to work well in this case, but the magnitude of this particular bias in Bangladesh is fairly small and it remain. to be seen how well the method will work in other application..

### *Beware of Model Life Table Specification*

Model life table families enter into the calculation of child survivorship estimates at several points, but certain calculations are more vulnerable than others to departures of the actual, unobserved pattern of mortality from the assumed model pattern. The use of proportions of deceased children to estimate  $q(x)$  values is of course independent of any model life table specification. The multipliers used to refine these estimate, do depend on model life table specification, but the multipliers are relatively insensitive to the model life table family chosen (Brass and Coale 1968:112—114). The real danger in model life table specification comes in the translation of the  $q(x)$  values to a common value and in the attempt to estimate trends. The Coale—Demeny (1966) regional model life tables provide a simple way to get a bearing on the errors involved. There are four model,, labeled “West,” “North,” “East,” and “South,” each of which may be used to translate any  $q(x)$  value into any other life table value. Suppose for example that we have estimated a  $q(5)$  value of 0.2 from child survivorship data. The West model life table with

$q(5) = 0.2$  has a  $q(1)$  value of 0.130. The North model with this value of  $q(5)$  has a  $q(1)$  value of 0.115, about 10 percent lower, and the East model with this value of  $q(5)$  has a  $q(1)$  value of 0.145, about 10 percent higher. These figures are obtained by interpolation in the model life tables given in Coale and Demeny (1966). Thus, if the translation of a  $q(5)$  to a  $q(1)$  value is based on a West model table, an error of about 10 percent will be incurred if the true pattern of mortality conforms to the North or East model. This example is broadly representative, but two points should be made. First, the translation to  $q(1)$  is particularly sensitive to model life table specification. Translation from  $q(10)$  or  $q(15)$  to  $q(5)$ , for example, incurs errors of about five percent. Second, translations over long age span. can give errors up to 20 percent and, in a few extreme cases, even larger errors. Thapa and Retherford (1982) have shown how extreme a range of trends can result from letting the model life table family vary arbitrarily.