Extended abstract
Estimating Adult Mortality through Surveys: An unbiased Method using Data on Survival of Relatives

Emmanuela Gakidou and Gary King

With substantial declines in child mortality over the past few decades, increasing importance is being placed on the measurement of adult mortality. While there have been substantial investments in the measurement of child mortality through survey and census programmes, there have been relatively less intensive efforts devoted to improvements in adult mortality and to its measurement. Complete vital registration systems are still far off for most low-income developing countries despite the fact that reliable baseline measures of adult mortality are needed for key programmes on the adult health agenda. Estimates derived from sibling survival histories in household surveys have not been used widely and are generally considered to be underestimates of true mortality. Currently applied methods have been shown to provide reliable mortality rates under two assumptions: 1) that there is independence in probabilities of death within families; that is that siblings all share the same probabilities of death; and 2) that probability of death is not related to family size. The effects of these assumptions on biasing the true mortality rates have not been explored in the literature.

In this paper, we develop a new method to estimate mortality rates of a population of interest using data on survival of relatives, from household surveys. Our proposed method does not rely on the assumption of independence. Using a simulated population in a simplified context we introduce our notation and show how the probability of death for a period of interest can be computed from data collected through a survey at the end of the period of interest. We then explore how the current method and our proposed method perform in estimating true mortality rates in a population under various scenarios of violation of the two assumptions of independence.

Methods / Framework for measurement

Let j (j = 1, . . . ,N) denote an index for an individual randomly selected (with equal probability and with replacement) from a population group of interest at time 1. Denote by B_j the number of siblings in the family of respondent j (including responding j) at the Beginning of the period (or "Born" into the group at time 1), S_j the number of siblings in the family of respondent j who Survive to time 2, and D_j the number who Die by time 2, so that $B_j = S_j + D_j$. The proportion of those who die in this family is the mortality rate, calculated as $M_j = D_j/B_j = (B_j - S_j)/B_j$.

We are interested in drawing a sample of survivors at time 2 to infer the mortality rate or other quantities from the full sample identified at time 1; we model this by imagining that the full sample is drawn at time 1 from the population but is then pruned via death to yield the time 2 population. That is, selecting only survivors from the random sample at time 1 is equivalent to a random sample at time 2 from the population of survivors. Using τ for quantities that are inestimable and π for those that are estimable (via simple averages) from survey data collected at time 2, denote the probability distribution (or histogram of the population) of the number of survivors:

$$\mathrm{P}(S_j = s) = \begin{cases} \tau & \text{for } s = 0 \\ (1-\tau)\pi_s & \text{for s=1,2,\ldots} \end{cases} \tag{1}$$

where $\pi_s \equiv \mathrm{P}(S_j | S_j > 0)$ is the conditional distribution of the number of surviving siblings among those at time 2, such that $\sum_s \pi_s = 1$ and τ is the (unobserved) probability of a time 1 family having *s* siblings surviving to time 2.

We define the family sizes at time 1 among families with *s* surviving siblings at time 2 as

$$\mathrm{P}(B_j = b | S_j = s) = \begin{cases} \tau_b & \text{for } b = 1, 2, \ldots \text{ and fixed } s = 0 \\ \pi_{b|s} & \text{for } b = s, s+1, s+2, \ldots \text{ and fixed } s = 1, 2, \ldots \end{cases} \tag{2}$$

where $\pi_{b|s} \equiv \Pr(B_j = b | S_j = s, S_j > 0)$ can be estimated by the histogram of family sizes of those observed at time 2. For any (fixed) number of surviving siblings *s* (2) represents a proper probability distribution and so for example $\sum_b \mathrm{P}(B_j = b | S_j = s) = 1$.

To compute the distribution of the mortality rate we first compute the joint distribution of *S* and *B*:

$$\begin{aligned} \mathrm{P}(S_j = s, B_j = b) &= \mathrm{P}(B_j = b | S_j = s)\mathrm{P}(S_j = s) && (3) \\ &= \begin{cases} \tau\tau_b & \text{for } s = 0 \\ (1-\tau)\pi_{bs} & \text{for } s = 1, \ldots, b \text{ for fixed } b = 1, 2, \ldots \end{cases} && (4) \end{aligned}$$

where $\pi_{bs} = \pi_{b|s}\pi_s$ is the joint probability of s and b conditional on $s > 0$. Then, since $M_j = (B_j - S_j)/B_j$, $S_j = B_j(1 - M_j)$. Hence, $\Pr(S_j, B_j) = \Pr(B_j(1 - M_j), B_j)$ and

$$\begin{aligned} \Pr(M_j = m) &= \sum_b \Pr(B_j(1 - M_j), B_j) && (5) \\ &= \begin{cases} \tau & \text{for } m = 1 \\ (1-\tau)\pi_m & \text{for } 0 \le m < 1 \end{cases} && (6) \end{aligned}$$

which is a discrete distribution and so the condition 0 ≤ m < 1 is, to be more precise, m = 0, 1/b, 2/b, . . . , (b − 1)/b.

We now define the quantity of interest. To do this in an informative way, we first define d_j as 1 if respondent j dies between time 1 and 2. Thus, the quantity of interest, the probability of death (or the proportion of those in the population who die) for all people in the interval from time 1 to time 2, is

$$q = \frac{1}{N}\sum_{j=1}^{N} d_j = \frac{\sum_{f=1}^{F} D_f}{\sum_{f=1}^{F} B_f} = \frac{1}{N}\sum_{j=1}^{N} M_j \qquad (7)$$

where the first expression is the standard definition, the second defines q for a sample with one respondent per family (f = 1, . . . , F), and the third is defined for the family mortality rate at the respondent level. This third definition will prove useful in our correction to samples drawn at time 2. The respondents randomly selected with equal probability from the population at time 1 each provide information about all family members or, in other words, family-level information about B, S, and M (e.g., $M_j = M_j^{'}$ for all j and j' that are members of the same family). Thus, we can view each draw of an individual equivalently as a draw of a family selected with probability proportional to B_j . For example, families with five siblings are represented in the population with five times the frequency, and thus have five times the sampling weight, as a family with one sibling.

We now turn to sampling at time 2, and introduce index i (i = 1, . . . , n) for respondents that have survived to time 2 and thus appear in the time 2 sample ($n \leq$ N). Sampling at time 2 generates two key problems. The first is that selecting respondents at time 2 with equal probability is equivalent to sampling families proportional to S_i rather than B_i. Fortunately, both quantities are known for all observations sampled, and so to return to the desired B_i weighting, we replace the simple average of M_j in the last expression in (7) with the weighted average of M_i, using weight $W_i = B_i/S_i$: $\sum_{i=1}^{n} M_i W_i / \sum_{i=1}^{n} W_i$, which would be equivalent except for the problem to which we now turn.

The second problem with the sample drawn at time 2 is that families with no survivors (S_i = 0) are not represented at all, and so we have no chance of weighting to recover the full information. To be more precise, the missing information is the total number of siblings in families with zero survivors, and it needs to be added to both the numerator and denominator of the weighted average since $Bi = Di$.

We factor this quantity into $N\tau\tau$, where $\bar{\tau} = \sum_b b\tau_b$ is the expected number of siblings in families without survivors. Since N is unobserved, we substitute: Using the fact that the time 2 sample size is $n = N - \tau N$, we solve for $N = n/(1 - \tau)$ and thus use $n\frac{\tau}{1-\tau}\bar{\tau}$ as our expression for the total number of siblings who died in families with zero survivors, leaving our estimator, conditional on τ and $\bar{\tau}$, as

$$\hat{q} = \frac{\sum_{i=1}^{n} M_i W_i + n\frac{\tau}{1-\tau}\bar{\tau}}{\sum_{i=1}^{n} W_i + n\frac{\tau}{1-\tau}\bar{\tau}} \qquad (8)$$

No certain or directly estimable information about the quantities τ and $\bar{\tau}$ exist in a sample drawn at time 2, but it turns out significant statistical information does appear to exist. We thus extrapolate information about these quantities from information in the sample.

To do this, we first compute the proportion of families with k survivors (for $k = 1, 2, \ldots$) and fit a model predicting this with k; we then use the same model to extrapolate these back to the value of τ given k = 0. Similarly, we compute the average number of siblings with k survivors ($k = 1, 2, \ldots$), fit a model that predicts this with k, and extrapolate back to $\bar{\tau}$ given $k = 0$. In the data we have examined, the fit of a model for $\bar{\tau}$ is very good, and we find if we set aside data for one (otherwise observed) value of k > 0, we can predict this with a high degree of accuracy. This procedure of course offers no guarantees, but in-sample empirical evidence makes us optimistic. The model for estimates of τ do not fit as well, but the variability in q, due to changes in estimates of τ and $\bar{\tau}$ within the range of what is empirically plausible based on the data, is relatively small.

Our ultimate estimator is then (8) with these estimates for τ and $\bar{\tau}$ substituted in. Standard errors or confidence intervals can be computed via bootstrapping.

Simulation

We simulate a population at time 1, under certain fertility assumptions. We then expose the population to probabilities of dying. At time 2 we take a survey (or a census) of the population and estimate the probability of dying, using only information available at time 2. We estimate two probabilities of death, the one outlined above using our proposed method and the currently used in the literature (deaths over population at the start of the interval, excluding the respondents from the denominator).

There are six parameters that are flexible in our simulated population:

1. fertility rate
2. average mortality rate
3. correlation of mortality rate with family size at time 1
4. correlation of mortality rate within families
5. distribution of siblings across age groups
6. sample size at time 2

This simulation environment allows us to test how the estimates derived from the two methods deviate from the truth as the parameters in the simulation change. The method currently employed by the literature assumes parameters (3) and (4) to be equal to 1, that is that mortality is not related to family size and that all members of the same family have the same probability of dying. This method also is based on the assumption that siblings are spaced close together, that is that most siblings in a family would fall into one broad (15 year) age group. The method we are proposing in this paper does not make any assumptions about parameters (3), (4) and (5). We expect our method to better recover the true probability of dying as we modify parameters (3), (4), and (5).

We also plan to test how the two methods perform under different fertility and mortality scenarios. We will explore fertility and mortality patterns currently observed in sub-Saharan Africa in countries with an advanced HIV/AIDS epidemic, as well as South-East Asia and Latin America. Finally, we will explore how the two methods perform when different sample sizes are drawn at time 2, ranging from samples that would be expected in a population survey to a full census.