Missing Children: Indirect estimation of child mortality from census microdata

Extended Abstract for Consideration for PAA 2012

Joshua R. Goldstein Ester González-Prieto
Jutta Gampe Frederick Peters

September 23, 2011

Abstract

In this paper, we develop a new method of inferring child mortality from the age gaps between surviving children. Based on the idea that higher child mortality produces an increased frequency of large gaps between surviving children, we use microsimulation to estimate the mortality rates implied by the observed distribution of age gaps. Application to populations with known child mortality shows that the method can reproduce well the estimates and differentials in child mortality seen in real populations. Estimates of child mortality from census-like data could be a valuable new addition to the toolkit of demographers to apply to the proliferation of historical census data via the IPUMS, NAPP, and the new Mosaic project, allowing the study of mortality and fertility (via the own child method) within populations and across time and space. This method, if successful, would enable researchers to estimate both mortality and fertility from a single census cross section, allowing full advantage of the richness of available census material.

1 Overview

Censuses represent the stocks of population counted at a particular moment in time. Just as the age-pyramid reflects the history of flows of births and deaths for the population, the ages of children within a family reflect the history of births and deaths within that family. Demographers have long used the "own-child method" in combination with an *external* estimate of mortality to estimate fertility rates. In this research, we develop methods to estimate child mortality from the observed age intervals between surviving children *internally* available within census microdata. This method, if successful, would enable researchers to estimate both mortality and fertility from a single census cross section, allowing full advantage of the richness of available census material. Research on social class differentials in fertility and mortality, regional differences, and change over time would all become possible in the many cases where census or census-like data is available.

The development of new methods to take advantage of the rich and growing availability of census microdata is an important opportunity for demographers. Building on the strong tradition of indirect methods in demography, the approach we take here is an example of what can be done by applying demographic thinking to microdata.

Some of the advantages of using census data for demographic estimation include:

- First, wide availability of census data makes it possible to compare across many populations, and over time. There is already considerable availability of historical material through the IPUMS and NAPP projects, and the Mosaic project promises to extend availability to historical Europe. There are even applications to ancient populations, thanks to the surviving census material from Roman Egypt, Ancient China, and Ancient Babylonia.

- Second, census material allows the study of differential demography because information on covariates is available. One can compare different regions, urban and rural areas, and different social classes (by occupation for example).

- Finally, as with many indirect methods, approaches that use the same source for events (numerators) and risk (denominators) are more robust, because data missingness will in effect cancel out when ratios are calculated. This is not the case when relying on separate birth and death recording, unless the errors happen to be of exactly the same size.

The key idea motivating this paper is that the ages of surviving children can reveal more than just how many children were born: they can also provide information on whether any children have died. As an example, consider the following two families from Alabama listed in the U.S. Census of 1850 (available at IPUMS-USA).

Table 1: Listing from the 1850 US Census

Name	Age
James W. Rhea	35
Julia Rhea	27
Amelia Rhea	7
Elizabeth Rhea	5
Jamey Rhea	3
Ephraim Knight	41
Mary James Knight	34
Permelia C. Knight	13
William N. Knight	11
Martha F. Knight	9
John A. Knight	1

In the Rhea family, we see that there are three children, each spaced 2 years apart. In the Knight family, there are four children – the first three are each two years apart but the fourth has a gap of 8 years. What could have caused such a long gap? It is possible that it simply could have taken a long time for John to have been conceived. Or, there could have been another child (or children) born in the 8 year period that died before they could be counted in the census.

Looking at just one family it is impossible to say whether a large age gap between surviving children indicates a child's death. Conception is a random process and can well lead to the occasional quite long interval. Furthermore, when children die very young, it can speed the time to the conception of the next child by ending lactation. In order to isolate the effect of mortality levels, our approach is to use microsimulation to take into effect the randomness of conception and the effect of lactation. Given an assumed fecundity level, life table, and post-partum insusceptible period, our micro-simulation generates distributions of surviving birth intervals. Our approach is to generate many combinations of these paramters until we find one that best fits the data. The result is then empirically based estimates of the life table, post-partum insusceptibility, and fecundity for our observed population.

In this extended abstract, we begin briefly by describing the nature of historical census data, noting the informative features that we are trying to incorporate into our estimate. We then decribe our micro-simulation model. Finally, we describe the estimation procedure and provide an example application to the 1850 data.

Before the PAA meetings, our plan is to (1) further refine the micro-simulation to include changing fecundity by age (2) to do sensitivity testing using simulation of heterogeneous fecundity across individuals (3) to validate the methods using a larger sample of populations with known mortality, and (4) apply the methods to an interesting substantive example, for example, differential infant mortality by class in the UK.

We have done some preliminary analysis of heterogeneous fecundity and found that the methods can also work well if appropriately modified. Our early validation experiments have provided reasonably accurate estimates of known infant mortality.

2 Data Requirements

The data required for our approach are listings of children's ages, by single year, by mother. For example, consider the listing from the 1850 US Census presented in Table 1. From such data we compute the age intervals between surviving children. (This can be done for the entire population, or for a sub-population of interest). We denote the fraction of the intervals of length i as P_i. The intervals for the United States in 1850 are shown in Table 2. We note that such distributions can be calculated for even small village-size sub-samples.

Table 2: Birth intervals of listing from the 1850 US Census given in Table 1

i	0	1	2	3	4	5	6	7	8
P_i	0.015	0.127	0.456	0.210	0.097	0.045	0.027	0.013	0.008

3 Micro-simulation and indirect estimation

Our micro-simulation of surviving birth intervals is governed by three parameters: (i) a baseline fecundity parameter p giving the monthly chance of a fully fecund woman to conceive (ii) ℓ that represents the length of the period of lowered fecundity due to the combined effects of post-partum and lactational amenorrhea after birth, and (iii) m that gives the mortality regime.

In our model, we consider only pregnancies leading to live births, with no explicit modelling of miscarriages or abortions. Under this assumption, at any time, a woman can be in one and only one of the following states:

- S_0: fecundable state when she is susceptible to pregnancy
- S_1: pregnant state when she is then completely insusceptible to a further pregnancy
- S_2: postpartum period after a live birth when she has lowered susceptibility to pregnancy.

We model the intervals between births as resulting from the allowable transition between those three states. Then, the length of the birth intervals will depend on the duration of the following transitions:

- T_1: Getting pregnant while fully susceptible ($S_0 \to S_1$)
- T_2: Giving birth ($S_1 \to S_2$)
- T_3: Becoming fully susceptible ($S_2 \to S_0$)
- T_4: Getting pregnant with lowered susceptibility ($S_2 \to S_0$)

A diagram of this model is shown in Figure 1

The monthly chance of conceiving a live birth is denoted by p if the woman is fully susceptible to conceive ($S_0 \to S_1$), and by p_L if she is in a lowered fecundity period ($S_2 \to S_1$). In current simulations, p_L is set to $p/10$ and the length of pregnancy ($S_1 \to S_2$) is fixed to 9 months.

The transition from lowered to full susceptibility ($S_2 \to S_0$) can occur either because the period of lowered susceptibility comes to an end, or because the child that is being nursed dies. In the micro-simulation a random number L is generated for the length of lactation conditional on the survival of the child and a random number X is generated for the age of death of the child.

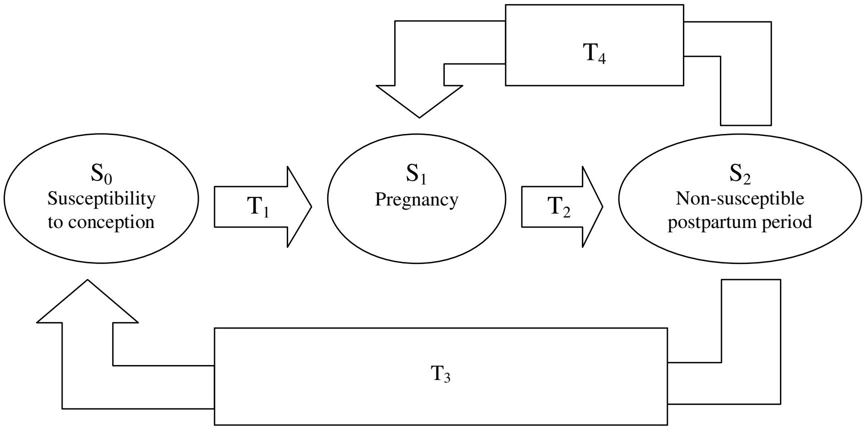

Figure 1: Possible states and transitions in our model

The transition ($S_2 \rightarrow S_0$) then occurs at time equal to the minimum of these two random values.

Our current approach is to use a common value of p for all potential birth intervals. Thus there is no aging of fecundity. There is also no heterogeneity across individuals. In future work, we will introduce heterogeneity.

Our current mortality modeling approach is based on picking the best fitting Coale-Demeny life table. We plan to investigate the use of simpler parametric approaches such as the Weibull or the Siler model of child mortality. An advantage of these parametric approaches is that they are easy to adapt to a more continuous search over parameter space, as opposed to the discrete grid search we now use.

One could also add more complexity to the model. For example, miscarriages could be included. Twinning could be included. Also, the post-partum period could be modeled in a more complex way including both post-partum ammenorea and lactational amenorea as separate factors.

With large data sets it is probably possible to fit many of the parameters based (for example) post-partum behavior or population heterogeneity based on the data. On the other hand, other factors such as miscarriage rates, could be included by taking standard values from the literature.

3.1 Estimation criteria and an example

In order to estimate which combination of mortality, breastfeeding pattern, and fecundity best matches our observed population, we first calculate simulated surviving birth intervals across a wide range of mortality schedules, lactation lengths, and baseline fecundity values.

For each triple (ℓ, m, p), we obtain a distribution of apparent birth intervals, denoted by $Q(i|\ell, m, p)$. Once we have generated a 'scenario grid' with all possible discrete values for the triple (ℓ, m, p), we choose the simulated distribution, $Q(i|\ell, m, p)$, that most resembles the observed apparent birth intervals distribution $P(i)$. For that purpose, we use the Kullback-Leibler divergence from the observed to the simulated distributions:

$$KL(P, Q|\ell, m, p) = \sum_i P(i) \log \left[\frac{P(i)}{Q(i|\ell, m, p)} \right] \tag{1}$$

The Kullback-Leibler is a (non-symmetric) measure of the difference between $P(\cdot)$ and $Q(\cdot|\ell, m, p)$, is always non-negative and it is equal to zero *iff* $P(\cdot)$ and $Q(\cdot|\ell, m, p)$ are identical.

The simulated distribution that most resembles the observed distribution is the one that minimizes $KL(P, Q|\ell, m, p)$. Then, the triple (ℓ^*, m^*, p^*) that minimizes $KL(P, Q|\ell, m, p)$ is chosen as the best set of estimated values.

As an example, we provide the simulated distributions $Q(.)$ for three example scenarios and for the best scenario that we found by searching the grid of parameter values. The first example sets infant mortality to 29/100 and a post-partum period of 1.0 years. The second example sets infant mortality lower at 23/100, and produces a slightly better fit as measured by a KL value of 0.09 instead of 0.10. The third example retains the lower infant mortality but lengthens lactation from 1 to 2 years worsening the fit considerably. The "best fite", which was chosen by looking over a broad range of parameter values including varying p, estimated the infant mortality to be about 140/1000 and lactation to be about 1.25 years. The fit is quite good, as can be seen by comparing it to the observed distribution and by the small KL value.

Table 3: Simulated distributions of intervals between surviving children for three example scenarios, the best fit, and observed in the 1850 US Census. The KL distance measures the similarity with the observed distribution.

	Example 1	Example 2	Example 3	Best Fit	Observed
Fecundity p	0.20	0.20	0.20	0.14	-
Mortality m	0.29	0.23	0.23	0.15	-
Lactation ℓ	1.00	1.00	2.00	1.25	-
$Q(0)$	0.00	0.00	0.00	0.00	0.02
$Q(1)$	0.23	0.27	0.03	0.13	0.13
$Q(2)$	0.32	0.33	0.26	0.48	0.46
$Q(3)$	0.16	0.15	0.33	0.18	0.21
$Q(4)$	0.11	0.12	0.15	0.10	0.10
$Q(5)$	0.08	0.06	0.10	0.06	0.05
$Q(6)$	0.05	0.04	0.08	0.03	0.03
$Q(7)$	0.04	0.02	0.05	0.02	0.01
KL	0.10	0.09	0.24	0.02	-

4 Extensions

Our research plan includes extensive validation and refinement of the micro-simulation model and the introduction of variable fecundity across individuals and over age. The method will be tested on a number of historical population as well as modern high-infant mortality populations such as those in the African Demographic and Health Surveys.

A further extension is to provide uncertainty intervals around our estimates. The micro-simulation approach allows us to do this in a straightforward way by repeated simulation and bootstrapping.

5 Discussion

In this research, we develop an approach to infer the likely pattern of mortality that generated observed intervals between surviving children. Such a method, if successful, would allow demographers to take advantage of the proliferation of available micro-data available in historical and modern censuses through the IPUMS, NAPP, and Mosaic projects. Our hope is that with such methods could enable a new generation of comparative analysis of child mortality and fertility, based on census micro-data.

Some drawbacks of the approach are (1) the problem is not analytically tractable and so requires a computer to perform estimates (2) the microsimulation is limited in applications to populations in which we believe we are able to model the determinants of birth spacing, so for example the use of contraception and/or parity-specific spacing behavior is not included in the model (3) randomness and errors stemming from small population and from age-misreporting can all introduce errors into the population. One can simulate the potential effects, but one still has to recognize that such errors will reduce the precision of the method.

On the other hand, despite its limitations, we believe there is great promise in this indirect estimation approach to estimating "missing children" from the age intervals between surviving children.

There are many possible refinements. One approach that might work, especially when there are many children per household, is to look for variation on a per-woman basis, instead of combining all women together into a common interval distribution, as is currently done. Relatedly, our approach could be used to "impute" missing children to individual women. The advantage of this approach would be that the individual records could then be used for standard statistical analysis, which via multiple imputation could provide appropriate standard errors.

The availability of historical census microdata is rapidly increasing but the demographic methods used for analysis tend to be the same as those developed long ago for aggregate data. One approach that may open new frontiers is to combine computationally intensive statistical estimation with rich micro-data. The use of micro-simulation of conception, birth, and survival given here is one such example.