DHS_U5M:

A flexible SAS macro to calculate childhood mortality estimates and standard errors from birth histories

By: Sidney Atwood, Brigham and Women Hospital Department of Global Health Equity
Dana Thomson, Harvard Medical School Department of Global Health and Social Medicine

Acknowledgements:
Adaptation of code by Keith Purvis, MeasureDHS
Adaption of partial month method by Kenneth Hill, Harvard School of Public Health

Date: December 2011

About:

This SAS macro generates childhood mortality estimates (neonatal, post-neonatal, infant ($_1q_0$), child ($_4q_1$) and under-five ($_5q_0$) mortality) and standard errors based on birth histories reported by women during a household survey. We have made the SAS macro flexible enough to accommodate a range of calculation specifications including multi-stage sampling frames, and simple random samples or censuses.

Childhood mortality rates are the component death probabilities of dying before a specific age. This SAS macro is based on a macro built by Keith Purvis at MeasureDHS. His method is described in *Estimating Sampling Errors of Means, Total Fertility, and Childhood Mortality Rates Using SAS* (www.measuredhs.com/pubs/pdf/OD17/OD17.pdf, section 4). Also see Appendix A for a description of this method. More information about Childhood Mortality Estimation can also be found in the *Guide to DHS Statistics* (www.measuredhs.com/pubs/pdf/DHSG1/Guide_DHS_Statistics.pdf, page 93).

We allow the user to specify whether childhood mortality calculations should be based on 5 or 10 years of birth histories, when the birth history window ends, and how to handle age of death with it is reported in whole months (rather than days). The user can also calculate mortality rates within sub-populations, and take account of a complex survey design (unequal probability and cluster samples). Finally, this SAS program is designed to read data in a number of different formats.

See below for a description of each macro input, and sample code. See Appendix B for an explanation of the partial month method used by Kenneth Hill.

Recommended citation: Atwood S, Thomson D. 2012. DHS_U5M: A flexible SAS macro to calculate childhood mortality estimates and standard errors from birth histories. Wrapper for: Purvis K. 1995. Mortality rates based on chapter 7 of the DHS report v.5.0.1. Retrieved month, year <www.websitename.com>.

SAS Macro Inputs:

Input Variable	Description	Default Value
DATA	**The dataset name** (and library location). Each record should represent one birth.	[Last used dataset]
PSU	**Primary Sampling Unit (PSU) ID** if the sample was a multistage cluster sample. PSU = V021 in the DHS. If there is no PSU, a bootstrap-like option is available. Specify PSU = bootci ###, where ### is the number of bootstraps you want to perform (this creates ### pseudo-PSU's).	V021
BYVAR	**Categorical variable describing sub-populations of interest**. The DHS commonly presents childhood mortality statistics by region (V024) and urban/rural residence (V025). If two or more variables describe your subpopulations (e.g. urban/rural populations within regions), then generate a new single variable to uniquely identify the smallest units in the sub-population. To ensure correct standard errors, do not delete any records (categorize all observations). There are no limits to the number of sub-populations that can be specified, though if the number of observations in a sub-population is small, standard errors will not be calculated.	[NONE]
DOB	**Child's date of birth (DOB)**. In the DHS, variable B3 represents DOB in century-month format. Depending on the format of DOB in your dataset, you may specify more than one variable: • CMC (century-month format): [variable] • SAS date: [variable] • mm yyyy: [month variable] [year variable] • dd mm yyyy: [day variable] [month variable] [year variable]	B3
AOD	**Age of death (AOD) in months**. This variable can be formatted as an integer or as a fraction. In the DHS, B7 represents age of death in months as an integer.	B7
FACTIONAL	This statement effects how the age of death value is used in the survival analysis. If AOD is an integer and DOB is in CMC format (as is the case for DHS datasets), you can **add a partial month, or random day, to AOD and DOB using the FRACTIONAL statement** thereby improving estimates of the probability of death within each age category. See Appendix B for a full explanation. If your dataset already has day, month, and year information for AOD and DOB, then age of death should be a fraction (e.g. 14.8 months). Specify one of the following values in the FRACTIONAL statement: • 0: Use the AOD variable as it is (an integer or fraction) • 1: add a uniform random partial month to AOD between 0 and 1 (if AOD is an integer) • 2: add a uniform random partial month to AOD between -0.5 and 0.5 (if AOD is an integer)	0

WEIGHT	**Sample weight for each observation**. In the DHS, V005 is the sample weight multiplied by 1,000,000 so that the value is stored as an integer. You may specify either V005 (integer) or the actual sample weight (fraction) here. If WEIGHT is not specified, then a weight of 1 is assigned to each observation.	V005
SURVEY_DATE	**Refers to the date of the survey with the mother**. SURVEY_DATE, SURVEY_DATE_FLAG, and YEARS are used together to determine which records from the birth history are used in the childhood mortality calculation. Depending on the format of your dataset, you may specify more than one variable here. Alternatively, you may specify a single date that marks the end of the birth history window for all women. If you specify SURVEY_DATE_FLAG = 1 or 0, and you have collected the survey date for each mother, then enter the following variables: • CMC (century-month format): [variable] • SAS date: [variable] • mm yyyy: [month var] [year var] • dd mm yyyy: [day var] [month var] [year var] If you'd like to specify the final date in the birth history window, enter that date here using the format described in SURVEY_DATE_FLAG.	V008

SURVEY_DATE_FLAG	Birth histories from the last 5 or 10 years are used to calculate childhood mortality. You can **specify when that birth history window ends:**
	• 0: Window ends at the date of interview of the mother
	• 1: Window ends at the date of first interview in the sample
	• If you specified a date in SURVEY_DATE, then enter the format of that date here. Examples:
	  ○ MMDDYY8.
	  ○ DDMMYY10.
	  ○ DATE10.
		0

YEARS	**The window of 5 or 10 years of birth histories to be included in the calculation**. The user may only specify 5 or 10 years.	5
AAI	**A binary variable that indicates whether the child was alive at the time of interview**. 0 indicates the child was dead, and 1 indicates the child was alive. In the DHS, this information is stored in variable B5.	B5
OUTLIB	**The name of the library to store the output**	WORK
count	This variable can be used to create interim output during the jackknifing process, but should be left blank unless you are troubleshooting a problem with the macro.	[NONE]

Sample code:

```
/*    DIRECTIONS marked in green: */

/*    1. store the DHS birth recode dataset in your working directory
      in a folder called "data". DHS data are produced in an old version of
      SAS (version 6). Any new files created must be stored in a modern (e.g.
      version 11) library */
      LibName dhs v6 "M:ChildMortality\Mali2006\data";
      LibName mort "M:ChildMortality\Mali2006\analysis";

/*    2. change the dataset name and library; and keep only the variables
      relevant to your analysis */
      data mort.br(keep = b: v005 v008 v021 v025);
      set dhs.MLBR52FL;
      run;

/*    3. load the macro */
      %include " M:ChildMortality\Mali2006\DHS_U5M.sas";

/*    4. run the macro */
      %DHS_U5M(DATA=br,
               PSU = V021,
               BYVAR = V025,
               DOB = B3,
               AOD = B7,
               WEIGHT = V005,
               SURVEY_DATE = V008,
               SURVEY_DATE_FLAG = 0,
               AAI=B5,
               FRACTIONAL = 0,
               YEARS = 5,
               OUTLIB = mort);
```

Appendix A. Explanation of MeasureDHS's Childhood Mortality Estimations Approach

Here is a description of the mortality estimation methods used in Keith Purvis's macro which we adopted and adapted:

> *The computation of these rates… follows the procedure developed by Rutsein (1984)[1]. Probabilities of death during a specific period of interest are obtained from probabilities calculated for 8 age intervals: less than 1 month, 1-2 months, 3-5 months, 6-11 months, 12-23 months, 24-35 months, 36-47 months, and 48-59 months. The probability of death for any age interval is defined as the ratio of the number of deaths occurring to children who were exposed to death in the age interval to the number of children exposed. The probability of death for a subinterval is obtained by subtracting the probability of surviving from 1. The… macro calculates the probability of surviving for each subinterval and the resulting probabilities are aggregated to obtain the mortality rate using the following formula:*
>
> $$(n)q(x) = 1 - \prod_{i=x}^{i=x+n} (1 - q(i))$$
>
> *where (n)q(x) is the probability of death between ages x and x+n, and q(i) is the subinterval probability of death.*
>
> *The… macro closely approximates the post-neonatal mortality rate by subtracting the neonatal mortality rate from the infant mortality rate.*
>
> *The sampling error of each mortality rate is calculated from the five or ten years preceding the surveying using the jackknife replication method…. For each rate, the number of cases used in the calculation corresponds to the number of children exposed to death between the age limits during the period of interest. For both neonatal and postneonatal rates, the number of cases is the minimum of the number of children exposed to death for the age intervals less than 1 months, 1-2 months, 3-5 months, and 6-11 months. For the child mortality rate, the number of cases is the minimum of the numbers of children exposed to death for the age intervals 12-23 months, 24-35 months, 36-47 months, and 48-59 months. For the under-five mortality rate, the number of cases is the minimum number of children exposed to death for all of the eight age intervals. The weighted number of cases is used to compute the sampling error under simple random sampling for each mortality rate using the following formula:*
>
> $$SE(rate) = \left[\frac{rate(1 - rate)}{n}\right]^{\frac{1}{2}}$$
>
> *in which n is the unweighted number of cases.*

[1] Rutsein, S.O. 1984. Infant and child mortality: Levels, trends, and demographic differentials. WFS Comparative Studies No. 43. Voorburg, Netherlands: International Statistical Institute.

Appendix B. Explanation of the partial month generator specified in the FRACTIONAL statement

In 2010, Kenneth Hill shared a Stata program on the Harvard Population Center DHS Wiki to calculate childhood mortality statistics. This program uses a random day generator to improve estimates of the probabilities that a child dies before the age of five.

Unlike the DHS, Hill calculates the probability of death within *each month* through the 59th month of life. Although Hill's approach makes better use of granular data, it is computationally intensive and there is evidence that mothers tend to report rounded ages of death in older children (Guide to DHS Statistics, 2006). This SAS macro uses the DHS approach, calculating probability of death in eight age groups (0, 1-2, 3-5, 6-11, 12-23, 24-35, 36-47, 48-59 months of age). The addition of a random partial month smooths deaths throughout a month and can improve estimates of mortality especially for younger children. Here is the explanation Hill provided with his Stata program in 2010:

> *Since there is some ambiguity about the month of death for data recorded as CMC for births and month at death for deaths (for example, the death of a child born in CMC 1206 at age 6 months could have occurred in CMC 1212 or 1213, different calendar years, depending on when in the month the birth occurred and where in the month of death the death occurred) a pseudo-random number between 0 and 1 is added to the CMC birth and another pseudo-random number between 0 and 1 to the age at death. The random number generator is "seeded" at the beginning of the routine to make sure that different runs give identical results (though differences could still occur if, for example, a covariate is used and the order of the covariate categories is changed). Exposure time in a particular month of age and 12-month period is based upon the proportion of a month that each child spends in each cell of the Lexis Diagram.*
>
> *For the neonatal interval, where more detail is available on age at death in the form of deaths by day, this information is used to better locate the month of death and also to locate the exposure time of the children who die in the first month. It is also used to calculate the average age at death in days, used in the conversion (below) of rates to probabilities.*
>
> *Once deaths and exposure for each Lexis cell have been summed for one 12-month period, the routine moves on to the next 12-month period; when it has completed all the requested 12-month periods, the events and exposure by month of age and 12-month period are summed for a given age, and event-exposure rates are calculated for each month of age*

$$ {}_1q_x = \frac{{}_1M_x}{\left(1 + 0.5 * {}_1M_x\right)} $$

> *Since in this case the unit of age is the month, the rates (and corresponding probabilities) are for months, and have to be converted into units of years to calculate standard indicators of probabilities of dying in childhood. Once again, standard life table methods are used to convert monthly probabilities to the standard indicators. Survivorship ratios are chained together to obtain probabilities of dying as follows:*

$$ {}_nq_x = 1 - \prod_{y=x}^{x+n-1}\left(1 - {}_1q_y\right) $$