Computing approaches to obtaining data for the indirect estimation of childhood mortality using Demographic and Health Survey data: A programming note

Sulaiman M. Bah.
Department of Statistics and Demography, University of Swaziland

To apply methods for the indirect estimation of childhood mortality, the data needed are the number of children ever born and child deaths, cross-classified by five-year age groups of women. The article discusses three approaches for extracting this information from the DHS data. The first approach involves writing a program in a high-level computer language; the second involves writing a program using a statistical package and the third involves a combination of the first two approaches. After discussing the problems related to the first two approaches, the article shows that the third approach can be accomplished in one minute on a "486"-computer, while the second approach would take about forty-five minutes to compute.

Introduction

Indirect estimation methods are the primary means of obtaining demographic measures for most of the developing countries that lack an adequately developed vital registration system. For the indirect estimation of childhood mortality, most of the methods developed are variants of the original method developed by William Brass. To apply these methods, female respondents are asked in censuses and surveys about the number of children ever born (CEB) and child deaths (CD). Sometimes the CD is broken down by sex to "sons dead" (SD) and "daughters dead" (DD). This information is classified per five-year age group of women (occasionally, by five-year marriage duration group). Using this basic information and the number of women in each age-group (NW), only a few computational steps are needed to arrive at estimates of childhood mortality (United Nations, 1983).

Several software packages are available for estimating childhood mortality. Prominent among these are MORTPAK-LITE and QFIVE, both user-friendly programs developed by the United Nations. Given the availability of the software packages, the only problem researchers face is how to obtain data on CEB, CD and NW, cross-classified by age-group of women.

Information on CEB, SD and DD is collected in all Demographic and Health Surveys (DHSs) and is available in machine-readable form. The worldwide DHS program, initiated in 1984 by IRD/Macro Systems Inc., is designed to collect data on fertility, family planning and maternal and child health. The surveys have been conducted in many developing countries and have been valuable in providing a series of useful health and demographic statistics. By 1996 several countries had already conducted their second round of DHSs. This article addresses the computing approaches considered for obtaining information on CEB, SD and DD from the Zimbabwean DHS (ZDHIS) data.

Computing approaches

Writing a program in a high-level computer language to extract the information needed from the data

This approach is straightforward. Besides the data, it needs a data dictionary that shows the particulars of the variables in the data file, their location, names, labels, values and value labels. This information comes with the DHS data, both in machine-readable and printed form. Given this data and the dictionary, all that is needed is proficiency in the writing of programs in one of the high-level computer languages like FORTRAN, PASCAL, C, etcetera. Unfortunately, even with these in place, one is still faced with other problems. One of these problems is that there are multiple records per case, at times more than ten. While this problem can be overcome, it calls for additional care in programming. Another problem, a more formidable one, is the size of the DHS files. They are quite large and take up large amounts of computer memory, often about 1 megabyte. This is too large to fit into the RAM available to some high level compilers, for example the FORTRAN compiler, WATFOR-77 (Coashi & Schueler, 1989; Kreitzberg & Carpenter, 1984). As such, this approach, although conceptually straightforward, is not always feasible.

Writing a program using a statistical package to extract the information needed from the data

Statistical packages such as SPSS/PC+, SAS, etc., could also be used to extract the information needed from the data. These packages are designed to handle very large data sets, and as such they can easily handle DHS data files.
Besides the data itself, these packages also need a data dictionary. To facilitate analysis of the DHS data with SPSS/PC+, the DHS data file is distributed with a special SPSS/PC+ file with the extension “.SPS”, which contains this dictionary. This file has to be appropriately edited to suit the researcher’s needs. The file is, however, still large and editing it is a time-consuming task as it still requires precise information about the structure of the data. To remove this burden, a special FORTRAN program, called SELECT, was designed by Charles Hammerslough in 1991. This program is a pre-processor for creating “INCLUDE” files of SPSS code (with the extension “.INC”). The INCLUDE file is then run under SPSS/PC+ to create a subset of the DHS rectangular file (according to the researcher’s needs) in SPSS format, including labels, value labels, and missing values. DHS data is now being distributed with the SELECT program.

The INCLUDE file can then be edited and, with a few commands, the required information can be obtained. There are at least two ways of obtaining the information needed. The simplest involves using the CROSSTABS command to produce cross-tabulation of CEB, SD, and/or DD by five-year age groups of women. The format is awkward however. The tables produced are for the frequencies of women reporting the different numbers of CEB, SD and DD. For example, for women in any given age group, the number of women reporting 0 CEB is given, then those reporting 1 CEB, 2 CEB, ..., 17 CEB. To obtain the CEB for women in any age group, the totals have to be computed manually. For example, for those women in the age group 15 to 19 years, the CEB is \((141 \times 1) + (24 \times 2) + (1 \times 3) = 192\). For the age group 20 to 24, the CEB is \((284 \times 1) + (182 \times 2) + (94 \times 3) + (31 \times 4) + (6 \times 5) + (1 \times 7) = 1091\).

Theoretically, this method can give the required data on CEB, SD and DD, but putting it into practice is labourious and involved, especially if the same calculation has to be repeated for various population subgroups. For example, if the infant and child mortality of various regions (say five regions) and the various levels of education of the women (say three levels) are to be compared, the same manual calculation would have to be repeated eight times each for CEB, SD and DD (leaving out the interactions). This is a very labourious and time-consuming task with high chances of error! One way of solving this problem is to transfer the output of the cross-tabulation to a spreadsheet package (e.g. Lotus 1-2-3 or Symphony) and then quickly do the required computations. This requires additional steps however.

Alternatively, the SPSS/PC+ can be programmed to perform the additions by using a combination of COUNT and COMPUTE statements (Norusis, 1990). The whole program is repetitious and can easily run into ten pages. The core of the program is the same. Although this method can give the desired results, it takes time to do the calculations. It took about forty-five minutes to run the program on a “486”-computer with a speed of 25 MHz.

**Using a combination of statistical packages and programming, using a high-level computer language**

Another approach is to use both the statistical packages and a high-level computer language. Among other things, statistical packages are ideal for extracting smaller data sets from very large data sets. The high-level languages, on the other hand, are excellent for making speedy computations on data sets of manageable size. In this approach, analysis could be done in two stages. In the first stage, a simple program (shown in the Appendix) is written in SPSS/PC+ to produce an ASCII output file with the relevant variables, i.e. CEB, SD and DD. This ASCII file then serves as an input file for a shorter FORTRAN program (also shown in the Appendix). The program gives not only the CEB, SD and DD but also CD (and its complement, CS, the total children surviving), which is the information actually required to serve as input in the programs for indirect estimation such as MORTPAK or QRTIVE. The program and the results are shown in the Appendix. The entire process took about one minute; about fifty-four seconds for the SPSS/PC+ program and about 6 seconds for the FORTRAN program.

**Conclusion**

Three approaches for extracting the information needed for the estimation of childhood mortality from the DHS data have been discussed. It is almost invariably assumed that a researcher using DHS data will apply the second approach, namely using SPSS/PC+ to analyse the data. However, because of the structure of the DHS data and the nature of the SPSS/PC+ package, it is problematic to use this approach to extract information on CEB, SD and DD classified by five-year age groups. A third approach is therefore recommended. This approach involves the use of both SPSS/PC+ and a FORTRAN program to produce the required output in a matter of seconds.

Finally, it is claimed that SPSS Windows offers some commands that can overcome this limitation of the SPSS/PC+. As the author did not have access to SPSS Windows at the time of writing, this claim could not be put to the test. It is recommended that researchers who do not have access to SPSS Windows and are using the SPSS/PC+ method at present, should adopt the faster approach as set out in this article.

**Note**

1. The research on which this article is based, was carried out while the author was holding a Population Council Post-Doctoral...
Fellowship at the University of Zimbabwe. The Fellowship support is gratefully acknowledged.

References


Appendix

SPSS/PC+ program

SET MORE=OFF
/RESULTS='C: \REGION.DAT'.
GET FILE = 'C: \REGION.SYS'.
WRITE VARIABLES = V013 V201 V206 V207/FORMAT=NUMBERED,
FIN.

FORTRAN program

C FORTRAN PROGRAM TO CALCULATE NCEB, SD, DO AND PD

INTEGER NCASE, NV013, NV201, NV206, NV207
OPEN (UNIT=5, FILE='C: \REGION.DAT')
OPEN (UNIT=6, FILE='C: \REGION.OUT')
NCEB1=0
NSD1=0
NDD1=0
NCEB2=0
NSD2=0
NDD2=0
NCEB3=0
NSD3=0
NDD3=0
NCEB4=0
NSD4=0
NDD4=0
NCEB5=0
NSD5=0
NDD5=0
NCEB6=0
NSD6=0
NDD6=0
NCEB7=0
NSD7=0
NDD7=0
N1=0
N2=0
N3=0
N4=0
N5=0
N6=0
N7=0
5 READ (5*, END=999) NCASE, NV013, NV201, NV206, NV207
10 IF (NV013 = EQ. 1) GOTO 100
20 IF (NV013 = EQ. 2) GOTO 200
30 IF (NV013 = EQ. 3) GOTO 300
40 IF (NV013 = EQ. 4) GOTO 400
50 IF (NV013 = EQ. 5) GOTO 500
60 IF (NV013 = EQ. 6) GOTO 600
70 IF (NV013 = EQ. 7) GOTO 700
100 N1=N1+1
    CALL TOTAL (NV201, NV206, NV207, NCEB1, NSD1, NDD1, NCD1, NCS1)
    GOTO 5
200 N2=N2+1
    CALL TOTAL (NV201, NV206, NV207, NCEB2, NSD2, NDD2, NCD2, NCS2)
FORTRAN output (unedited)

<table>
<thead>
<tr>
<th>MOTHERS AGE GROUP (I)</th>
<th>NUMBER OF CEB WOMEN</th>
<th>SD</th>
<th>DD</th>
<th>TCD</th>
<th>TCS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>192</td>
<td>10</td>
<td>14</td>
<td>178</td>
</tr>
<tr>
<td>1</td>
<td>1 021</td>
<td>39</td>
<td>73</td>
<td>1 018</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>840</td>
<td>93</td>
<td>79</td>
<td>1 793</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>679</td>
<td>1 965</td>
<td>125</td>
<td>2 351</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>589</td>
<td>1 965</td>
<td>158</td>
<td>2 310</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>464</td>
<td>2 035</td>
<td>116</td>
<td>1 810</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>318</td>
<td>2 035</td>
<td>116</td>
<td>1 810</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>290</td>
<td>1 993</td>
<td>138</td>
<td>1 733</td>
<td></td>
</tr>
</tbody>
</table>