Max-Planck-Institut für demografische Forschung
Max Planck Institute for Demographic Research
Konrad-Zuse-Strasse 1 · D-18057 Rostock · GERMANY
Tel +49 (0) 3 81 20 81 - 0; Fax +49 (0) 3 81 20 81 - 202;
http://www.demogr.mpg.de

MPIDR TECHNICAL REPORT 2012-002
APRIL 2012

# An Excel spreadsheet for the decomposition of a difference between two values of an aggregate demographic measure by stepwise replacement running from young to old ages

Evgeny M. Andreev
Vladimir M. Shkolnikov (shkolnikov@demogr.mpg.de)

For additional material see www.demogr.mpg.de/tr/

# An Excel spreadsheet for the decomposition of a difference between two values of an aggregate demographic measure by stepwise replacement running from young to old ages

*by Evgeny M. Andreev and Vladimir M. Shkolnikov*


## *Abstract*

A general algorithm for the decomposition of differences between two values of an aggregate demographic measure of age and other dimensions is realized as Excel/VBA. It assumes that the aggregate measure is computed from similar matrices of discrete demographic data for two populations under comparison. The algorithm estimates the effects of replacement for each elementary cell of one matrix by the respective cell of another matrix. The replacement runs from young to old ages.


## *Background*

Comparative analyses often look at the differences between the values of aggregate demographic measures calculated for two populations. Once a difference has been identified, the researcher may be interested in decomposing it according to the effects of underlying factors, such as age group, sex, and cause of death.

The general decomposition problem is to estimate the additive contributions of the differences between the values of the factors to the overall difference between the values of the aggregate measure. The aggregate measure of interest is considered as a dependent function of the factors.

In 2002, we proposed a general algorithm for solving the decomposition problem by means of stepwise replacement (Andreev, Shkolnikov, Begun, 2002). For some aggregate measures, this algorithm produces formulae for the decomposition (Andreev, 1982; Pollard, 1982; Arriaga, 1984; Pressat, 1985; Das Gupta, 1994; Das Gupta, 1999; Shkolnikov, Andreev, and Begun, 2001, 2003).

In many cases, however, a direct numerical application of the replacement algorithm has to be performed. For example, the existing formula for the decomposition of a difference between two life expectancies by ages and causes of death degenerates if at one or more ages the all-cause death rates in the two populations are exactly the same. For many aggregate measures, the analytical expression for the decomposition either cannot be deduced, or it can be deduced, but is too complicated. For example, the formal expression of the age decomposition of a difference between e-dagger values (Vaupel, Canudas Romo, 2003) appears to be very complex (Shkolnikov, Andreev, Zhang, Vaupel, 2011).

The numerical algorithm for decomposition by stepwise replacements is described in detail (Andreev, Shkolnikov, Begun, 2002). The “Decomposition_replacement.xls” spreadsheet produces exactly the algorithm that was described in the article. The replacement runs from young to old ages, and in each age group, the program performs all of the possible sequences of replacements.

## *How to use it*

The Decomposition_replacement.xls spreadsheet consists of four worksheets: “C,” “Data-1,” “Data-2,” “Result,” and the VBA program “Decomp.”

Specify the number of age groups and the number of variables at each age in cells B2 and B3 of worksheet C, respectively. The total number of age groups (rows) should be less than 120, and the number of variables at each age should be less than 11.

In the Data-1 and Data-2 worksheets, place the data elements to be replaced in the cell ranges beginning from the B2 cells. The data elements for the youngest age should be placed in row 2, the data elements for the second-youngest age should be placed in row 3, etc. Column A and the row can be used for the column and the row titles.

In the Data-1 and Data-2 worksheets, place the formulae/table(s) for the calculation of the aggregate demographic measure in populations 1 and 2 from the data elements described in the previous paragraph. Make sure that all of the calculations in the cells are correct. Calculation errors can result from division by zero or negative arguments of the Excel functions LOG() and SQRT(), and for other reasons.

In cell B6 of worksheet C, place a formula for the calculation of the difference between the two values of the aggregate demographic measure. These two values are calculated in the Data-2 and Data-1 worksheets.

Run the program for the decomposition by clicking the "Run macro" button (worksheet C). It is possible to interrupt the execution of the program using Ctrl-Break, but this can lead to the loss of some initial data. If an error code (#NULL!, #DIV/0!, #VALUE!, #REF!, #NAME?, #NUM!, #N/A) appears in cell B6, it is a sign that the program has been interrupted and that some of the initial data may have been lost.

The decomposition results will appear in the Result worksheet.

The application of the program is illustrated using four examples.

### *Examples*

Each example is placed in the Excel file named *Decomposition_replacement_from_young_to_old_ages(K).xls*, where *K* is a number of the example. The examples illustrate the principles of using the spreadsheet program. They can also be used as outlined here, because each example provides a solution to a real research problem.

*Example 1. The decomposition of a difference between two life expectancies at birth by age and causes of death*

Problem

Age- and cause-specific death rates for men in the USA and in England and Wales in 2002 are given. Life expectancy at birth for the two populations is calculated from the death rates. The age groups are 0, 1-4, 5-9, …, 85+. There are seven causes of death: neoplasms, diseases of the circulatory system, diseases of the respiratory system, diseases of the digestive system, accidents and violence, and all other causes of death.
Data source: The WHO Mortality Database http://www.who.int/whosis/mort/download/en/

The difference between the life expectancies at birth between England and Wales and the USA in 2002 is to be split by age and cause of death.

Solution

The Data-1 and Data-2 worksheets are organized in the following way.

The range of age-cause-specific death rates is placed in the range B2:G20. The rows correspond to age groups, and the columns correspond to causes of death. An abridged life table (Chiang, 1984) is placed in the range L2:T20. Column L2:L20 contains a vector of age-specific death rates for all of the causes combined. These rates are calculated as sums of rows of age-specific death rates.

The expression for the difference between the life expectancies in population 2 and population 1 in cell B6 of worksheet C is ='Data-2'!T2-'Data-1'!T2, because the values of life expectancy at birth are located in cells T2 of the Data-1 and Data-2 worksheets.

An elementary step of replacement for age *x* and cause *c* includes the calculation of the life expectancy change resulting from the replacement of the death rate M(*x*, *c*, pop1) by M(*x*, *c*, pop2). This replacement is being performed for each combination of causes of death other than *c*, and the final effect is computed by averaging the resulting $2^{n-1}$ effects (for *n* causes of death).

With respect to age, the replacement progresses from the first age 0 to the last age 85+. The replacement of the entire range of ages runs twice. Pop 1 is being replaced by pop 2, and then pop 2 is being replaced by pop 1. The final (symmetrical) components are calculated by the averaging of the components resulting from the two replacement cycles (Andreev, Shkolnikov, Begun, 2002).

*Example 2. The decomposition of a difference between two values of lifetime losses ($e^\dagger$ e-dagger) by age and causes of death*

Problem

Data from the first example are used. The lifetime losses quantity ($e^\dagger$) for the two populations is calculated from age- and cause-specific death rates.

Data source: The WHO Mortality Database http://www.who.int/whosis/mort/download/en/

The difference between the $e^\dagger$ values for men in England and Wales and in the USA in 2002 should be split by age and causes of death.

Solution

The Data-1 and Data-2 worksheets are nearly the same as in the first example. Two new columns, U2:U20 and V2:V20, are added to the life table for the calculation of $e^\dagger$.

These two columns allow for the computation of $e^\dagger$ according to the following equation:

$e^\dagger = \sum_x \left[ d_x \frac{1}{2}(e_x + e_{x+1}) \right]$, where ${}_1d_x$ and $e_x$ are the life table deaths within age group [*x*, *x*+1) and life expectancies at age *x*.

The expression for the difference between the life expectancies in population 2 and population 1 in cell B6 of worksheet C is ='Data-2'!V2-'Data-1'!V2.

The sequence of replacement across the matrices *M*(*x*, *c*, pop i) is exactly the same as in the first example.

More information about the lifetime losses, time trends, and inter-country differences in the lifetime losses can be found in Shkolnikov, Andreev, Zhang and Vaupel, 2011.

*Example 3. The decomposition of a difference between two modal ages at death by age*

Problem

Age-specific death rates for women in Sweden in 2010 and 2005 are given. The modal age at death for the years 2005 and 2010 is calculated from these rates. Age groups: 0, 1, 2, …, 110+.
Data source: The Human Mortality Database www.mortality.org

A difference between the modal ages at death for Swedish women in 2008 and 2009 is to be split by age.

Solution

The Data-1 and Data-2 worksheets are organized in the following way.

The vector of age-cause-specific death rates ($M_x$) is placed in the range B2:B112. A complete life table (Chiang, 1984) is computed from $M_x$ values in the range E2:K112. The modal age at death is being calculated as the age at which the $d(x)$ function (life table deaths) reaches its maximum using the standard Excel functions MAX and VLOOKUP (cells M2:N4 and range L2:L112).

Note that unstable $M_x$ values at ages 106 to 110+ are replaced by $M_x$ at age [105, 106). Note, too, that our calculation uses empirical $M_x$ values. In many cases, it might be better to use smoothed $M_x$ values instead of (somewhat shaky) empirical ones.

The expression for the difference between the life expectancies in population 2 and population 1 in cell B6 of worksheet C is ='Data-2'!N6-'Data-1'!N6.

The replacement across the vector B2:B112 progresses from the first age 0 to the last age 110+. The full cycle that includes the entire range of ages runs twice (pop 2 replaces pop 1, and pop 1 replaces pop 2).

More information about the modal age at death can be found in the study by Shkolnikov, Andreev, Zhang, Vaupel, 2011.

*Example 4. The decomposition of a difference between two life expectancies at age 30 by education-specific mortality and population educational compositio.*

Problem

Life expectancy at age 30 for populations 1 and 2 is calculated from age-specific death rates that are computed as weighted sums of age-education-specific death rates and age-education-specific population weights. A set of the latter weights is referred to as educational composition. The age-education-specific death rates and the educational composition are given. The age groups are 30-34, 35-39, …, 85+. The educational groups are high (tertiary) education, medium (secondary) education, and low (lower than secondary) education. The educational composition is given as a set of three-element rows of age-group-specific weights. The data are given for two time periods: 1971-1975 and 1996-2000.

The difference between life expectancies at age 30 in 1996-2000 and 1971-1975 is to be split by age and (within each age group) by mortality in each of the educational groups, and by population composition.

Solution

The Data-1 and Data-2 worksheets are organized in the following way.

The elements to be replaced are in the range B2:E13. Within this range, the first three columns contain age-specific death rates for men with high (B2:B13), medium (C2:C13), and low (D2:D13) education; and the fourth column contains a categorical variable, “Educational structure,” that equals 1 or 2 for educational compositions of population 1 or 2, respectively.

The expression for the difference between the life expectancies in population 2 and population 1 in cell B6 of worksheet C is ='Data-2'!R2-'Data-1'!R2.

The replacement of one age group (row) includes replacements of three education-specific death rates and of the Educational structure variable. Each of these four replacements must be done $2^3$ times (all possible combinations of pop 1 and pop 2 values of the remaining three variables). The replacement progresses from the first age 30-34 to the last age 85+. The full

cycle across the entire range of ages runs twice (pop 2 replaces pop 1, and pop 1 replaces pop 2).

*Example 5. The decomposition of a difference between two total fertility rates by age and birth order*

Problem

Conditional age- and birth order-specific fertility rates (fertility rates of the first kind) for Sweden for the years 2000 and 2010 are given. The ages are 12 and younger, 13, 14, …., 54, 55+. The birth orders are 1, 2, 3, 4, 5+. The period total fertility rates are computed as a function of conditional age-order-specific fertility rates.

Data source: The fertility rates originate from data on annual births by age of the mother and birth order, and from the population of register-based females by age and parity. The conditional fertility rates are extracted from the Human Fertility Database at www.humanfertility.org.

A difference between the total fertility rates in 2000 and 2010 in Sweden is to be split by age and birth order.

Solution

The Data-1 and Data-2 worksheets are organized in the following way.

The conditional fertility rates to be replaced are located in the B2:F45 range. They are used for the calculation of the fertility table. This is an increment-decrement table expressing the progression of a population of 10,000 women from parity 0 to parity 4+ and age. The calculation of the table population of women by age and parity is in the K2:O45 range. The calculation of the table probabilities of giving $i$-th birth by a woman with $i$-1 children is in the R2:V45 range. The calculation of the table births by age of the mother and birth order is in the Y2:AD45 range.

The expression for the difference between the life expectancies in population 2 and population 1 in cell B6 of worksheet C is ='Data-2'!Y48-'Data-1'!Y48.

An elementary replacement of one age-order-specific fertility rate $f(x, i,$ pop 1) by fertility rate $f(x, i,$ pop 2). At each age, five elementary replacements are executed for $i$=1, 2, 3, 4, 5+. Each of them is performed $2^4$ times with all possible combinations of $f(x, j,$ pop 1) and $f(x, j,$ pop 2) for all values of $j$, such that $j \neq i$. The replacement progresses from the first age 12 and younger to the last age 55 and older. The full cycle across the entire range of ages runs twice (pop 2 replaces pop 1, and then pop 1 replaces pop 2).

More information about the compilation of the period fertility table and of the period total fertility rates can be found in Jasilioniene et al. (2011). Although our spreadsheet uses a slightly simplified version of these formulae for constructing the period fertility table, the difference hardly affects the TFR estimates.

## *References*

Andreev, E.M. (1982). Metod komponent v analize prodoljitelnosty zjizni. [The method of components in the analysis of length of life]. *Vestnik Statistiki*, 9, 42-47.

Andreev, E.M., V.M. Shkolnikov, and A.Z. Begun. (2002). Algorithm for decomposition of differences between aggregate demographic measures and its application to life expectancies,

healthy life expectancies, parity-progression ratios and total fertility rates. *Demographic Research* 7, 499–522.

Arriaga, E. (1984). Measuring and explaining the change in life expectancies. *Demography* 21(1), 83-96.

Canudas-Romo V. (2010). Three measures of longevity: time trends and record values. *Demography*. 47(2): 299-312.

Chiang, C.L. (1984). The life table and its applications. Malabar, Florida: Robert E. Krieger Publishing Company.

Jasilioniene A., Jdanov D.A., Sobotka T., Andreev E.M., Zeman K. and V. M. Shkolnikov with contributions from J.Goldstein, E. J.Nash, D.Philipov, G.Rodriguez. (2011). *Methods Protocol for the Human Fertility Database*, pp. 50-51

Pollard, J.H. (1982). The expectation of life and its relationship to mortality. *Journal of the Institute of Actuaries*, 109, Part II, No 442, 225-240.

Pressat, R. (1985). Contribution des écarts de mortalité par âge à la différance des vies moyennes. *Population*, 4-5, 766-770.

Shkolnikov, V., Valkonen, T., Begun, A., Andreev, E. (2001). Measuring inter-group inequalities in length of life. *Genus*, LVII(3-4), 33-62.

Shkolnikov, V.M., Andreev, E. M., Begun, A. Z. (2003) Gini coefficient as a life table function: computation from discrete data, decomposition of differences and empirical examples. *Demographic Research*, 8:11, 305-358

Shkolnikov V.M., Andreev E.M., Zhang Z., Oeppen J., J.W.Vaupel. 2011. Losses of expected lifetime in the United States and other developed countries: methods and empirical analyses. *Demography*, 48: 211-239

Vaupel, J. W., & Canudas Romo, V. with (2003). Decomposing change in life expectancy: A bouquet of formulas in honor of Nathan Keyfitz's 90th birthday. *Demography*, 40, 201–216.