Variability in Life-Table Estimates Based on the National Health Interview Survey Linked Mortality Files

Van Parsons[1], Nathaniel Schenker,
Kimberly Lochner, Gloria Wheatcroft and Elsie Pamuk
[1]National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Rd Rm 3219, Hyattsville MD 20782

Abstract[a]

This study investigates methods for assessing variability of estimated life-table functions when data are longitudinal and involve a complex survey. Traditional methods for estimating such variability, developed by Chiang, are appropriate when the observations are independent. With longitudinal data and a complex survey design, however, there can be two types of dependence: (1) within age groups, due to, e.g., clustering in the survey; and (2) across age groups, due to observations for a single individual contributing information for more than one age group. For complex survey data Chiang's method may ignore several components that contribute to estimator variability. As an alternative variance estimation procedure, we propose using a balanced repeated replication method. In the context of a study of disparities in life expectancy, for which data are obtained through linkage of respondents in the National Health Interview Survey to death records, we explore the alternative methodology.

Key Words: Balanced Repeated Replication, Variance Estimation

1. Introduction

Disparities in life expectancy by gender, race/ethnicity, education and income are of interest to health researchers and policymakers. While there are comprehensive data for life expectancy based on sex, age, and race (white and black) from the U.S Decennial Life Tables, data resources with high quality socioeconomic status (SES) indicators and mortality are limited. Cohort data that link individual records, with self-reported information on education and income, to mortality data can overcome this data deficiency. Currently, two data sources, the National Longitudinal Mortality Study (NLMS) and the NHIS Linked Mortality files, provide the most recent and comprehensive prospective mortality data for the United States. Using NLMS data, estimates of life expectancy by income for whites and blacks were published in Health, United States, 1998 (Pamuk, Makuc et al. 1998) and an updated examination of life expectancy by SES was published in 2003 (Lin, Rogot et al. 2003). However, the most recent versions of the NHIS Linked Mortality files present several advantages for this

[a] *The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.*

kind of analysis, e.g. the data files are of nationally representative U.S. cohorts, provide a large number of deaths, and allow for the examination of cause-specific deaths as well as time trends.

Chiang (1984) provides a comprehensive discussion of the operating characteristics of the standard life table. The estimation of a life table from complex survey data is somewhat straightforward, since the life table is based on component cell totals and their sums and ratios. While the estimation of the life table itself can be established rather easily, by replacing the required cell totals with corresponding survey weighted totals, the estimation of the standard errors of life expectancy is more complicated. Chiang has provided variance estimates, but under a much simpler sampling distribution than can be assumed by the combined years of the complex survey, NHIS. Previous papers using complex survey data for life tables have made reference to direct usage Chiang's variance estimation error methods or used partial design information in the computation of variances. In this paper we propose using a design-based balanced repeated replication variance estimator for the life tables based on a complex design. We feel that this method of variance estimation more accurately captures the stochastic nature of the complex design than those methods proposed in recent work.

This work was presented as a poster at the 2009 Joint Statistical Meetings. At the time of this writing, a comparison of approaches to variance estimation is being performed and is not complete. Consequently, this paper will only touch on the ideas of our strategy. We expect a complete paper to be submitted for publication in the near future.

2. Data Sources and Structures

2.1 National Health Interview Survey (NHIS) Linked Mortality Files

The National Center for Health Statistics (NCHS) conducts mortality follow-up for eligible NHIS participants through probabilistic record linkage to the National Death Index (NDI), which maintains a national file of death certificate records collected from state vital statistics offices. (For details see the National Center for Health Statistics, National Death Index URL: www.cdc.gov/nchs/ndi.htm.) We combined 11 years of the NHIS (1990-2000) that had mortality follow-up for eligible participants from the time of their interview through December 31, 2002, yielding about 800,000 eligible NHIS records between 1990 and 2000. We utilized the restricted versions of the files so as to have more complete information on age, interview date, and death date. At the time of this analysis the updated NHIS mortality follow-up, which is currently available, was not complete. For details see www.cdc.gov/nchs/r&d/nchs_datalinkage/nhis_data_linkage_mortality_activities.htm.

2.2 NHIS Complex Design

The NHIS design is documented in Massey et al. (1989) and Botman et al. (2000). The NHIS samples typically include about 30,000-40,000 households per year (90,000-100,000 persons). The major design features which should be considered in any complex analysis are

1. Households are clustered by geography (counties and Census-defined block clusters).

2. The sample is clustered over time; the same geographical units are revisited each year.

3. There is an oversampling strategy for targeted minority groups.

4. The years 1986-1994 and 1995-2004 cover two cycles of the NHIS, and the sampling between those two cycles is independently implemented.

5. The survey weighting procedures are rather involved. Non-response and poststratification weighting adjustments are implemented quarterly.

3. Life Table computations

3.1 Basics

Chiang (1984) provides the forms of the components needed to compute a standard life table. While some life tables are partitioned into one-year age groups, those based on NHIS 11-year survey database are often coarsened to 10-year intervals for reliable estimation over the broad range of subdomains considered. An abridged life table is derived from the fundamental elements of the following table.

Abridged Life Table Components for Calculation

Age interval, $[x_i, x_{i+1})$	D_i	P_i	M_i	q_i	a_i	e_i
[25, 35)						
[35, 45)						
[45, 55)						
[55, 65)						
[65, 75)						
[75, 85)						
85+				1.00		

From the data we must estimate:

D_i = number of deaths in interval i and

P_i = number of person-years in interval i

to produce the estimates for the key life table parameters,

$M_i = D_i \ / P_i$ = death rate in interval i and

$q_i = (n_i M_i) / [1 + (1 - a_i) n_i M_i]$

= probability of dying in interval i given survival to the beginning of the interval,

where n_i is the length of interval i and a_i is the average fraction of n_i lived by those who die within the interval (here assumed to be .5 for every interval except the last).

Using standard relations of the components of the above table, e.g., Chiang (1984), we can obtain an estimator of e_i = life expectancy at age x_i.

3.2 Estimating the Variance of an Estimated Life Expectancy: Chiang and BRR methods

Under an assumption that grouped data in the age intervals are *independent across intervals* and total deaths, D_i, within interval i can be modeled as having *a binomial distribution with probability* q_i but with unknown number of independent trials,

Chiang has developed Taylor-Linearization methods to estimate $V(\hat{e}_\alpha)$, where a "hat" denotes an estimated quantity:

$$\hat{V}(\hat{e}_\alpha) = \sum_{i=\alpha}^{w-1} \hat{p}_{\alpha i}^2 [(1-a_i)n_i + \hat{e}_{i+1}]^2 \hat{V}(\hat{p}_i), \qquad \text{(C)}$$

where w denotes the final age interval, $p_{\alpha i}$ is the probability of surviving to age x_i given survival to age x_α, and

$$\hat{V}(\hat{p}_i) = \hat{q}_i^2 (1-\hat{q}_i)/D_i \text{ (with } q_i = 1-p_i).$$

For survey data the sampling structures are different from those assumed in the derivation of Chiang's formulas.

For survey data we record the key variables

Subject	Survey year	Survey weight	Age at Interview	Age at Death or Censoring	Status Dead (1) Censored(0)
j	y	wt	t_0	t_d	*Status*

and then obtain the estimates

$$\hat{D}_i = \Sigma_y \, \Sigma_{j \in y} \, wt_j \cdot I(t_{j,d} \text{ in interval } i) \cdot Status_i \text{ and}$$

$$\hat{P}_i = \Sigma_y \, \Sigma_{j \in y} \, wt_j \cdot \text{length}\,([t_{j,0}, t_{j,d}] \cap [x_i, x_{i+1}])$$

where $I(t) = 1$ (0) if t is true (false).

While the weighted forms $\hat{D}_i$ and $\hat{P}_i$ can be used as the component estimates for the abridged life table, the stated assumptions necessary to use expression (C) are considered invalid, and other variance computation methods should be used.

We use a *Balanced Repeated Replication* (BRR) approach for estimating variances with the NHIS linked data. Four main reasons for selecting this approach were:

1. The variability resulting from the quarterly poststratification could be taken into account.
2. As a replicate method, the BRR procedure can be used to define a reproducible set of replicates as a function of a specific Hadamard matrix. Bootstrap and Jackknife procedures require a random seed which make complete reproducibility more difficult.
3. A set of BRR replicates could be used for other statistics using the NHIS Linked Mortality files. Furthermore, given a set of replicates one can compute variance estimates for complex survey analysis without resorting to specialized software.
4. A complete linearization approach to variance estimation of the life expectation estimator is somewhat complicated.

For variance estimation we set up a "*2-PSUs-per-Stratum"* pseudo-structure for the NHIS that captured much of the clustering of the NHIS.

NHIS Design years	Number of Strata	Number of PSUs
1986-1994	189	378
1995-2005	339	678

Lumley's *R* package *survey* **(**R Development Core Team (2009)**)** was used to compute 544 sets of replicate weights using a quarterly poststratification adjustment. A Fay-adjustment (Judkins(1990)) of 0.30 was used. Once created, the replicate weights can be used to estimate the variances of life table components.

4. Evaluations

4.1 Basic Life Table Estimates.

Survey-based estimates of life-expectancy could be subject to biases. As a coarse check we compared the point estimates of life expectancy derived from the NHIS to those produced by the U.S. Decennial Life Tables for 1999-2001for several select large domains that are common to both data systems. The results can be seen in Table 1. As the NHIS only targets the civilian non-institutionalized population, as expected the NHIS Linked Mortality overestimates life expectancy, but these comparisons suggest no gross biases.

Table 1 NHIS-NDI linked mortality files compared to U.S. Decennial life tables, 2000 (DVS)

	LE age 25			LE age 45			LE age 65		
	NHIS-NDI	(SE)	DVS	NHIS-NDI	(SE)	DVS	NHIS-NDI	(SE)	DVS
WM	51.3	(0.09)	51.0	32.6	(0.08)	32.5	16.3	(0.06)	16.2
WF	55.9	(0.08)	55.8	36.9	(0.08)	36.6	19.7	(0.06)	19.2
BM	46.5	(0.23)	45.4	29.2	(0.20)	27.9	14.9	(0.17)	14.1
BF	51.9	(0.21)	51.7	33.6	(0.19)	33.3	18.2	(0.16)	17.7

Life Expectancy (**LE**) by sex (**M**ale, **F**emale) for **W**hites and **B**lacks. For the NHIS-NDI, Whites and Blacks are of non-Hispanic origin.
SE = standard error based upon 544 replicate sample weights

In Figure 1 and Figure 2 we plot, respectively, life expectancy and its Relative Standard Error (RSE) for sex by income domains. In these figures income is ordered 1=Low to 4=High.

4.1 Ongoing Comparisons of Variance Estimation Methods

The past literature on Life Tables based on complex-surveys seems to focus on variance estimation methods adapted to Chiang's formula and at most used only partial complex-design information. As of this writing we are currently evaluating our BRR method in comparison to Chiang's method and adaptations. We expect a detailed evaluation to be submitted for future publication.

References

Botman S, Moore T, Moriarity C, and Parsons V. (2000), Design and Estimation for the National Health Interview Survey, 1995-2004, *Vital Health Stat* 2(130).

Chiang, CL (1984), *The Life Table and Its Applications*, Malabar, FL: Krieger.

Judkins, D (1990), Fay's Method for Variance Estimation. *Journal of Official Statistics*, **6**, 223-240.

Lin, CC, Rogot E , et al. (2003), A further study of life expectancy by socioeconomic factors in the National Longitudinal Mortality Study, *Ethn Dis* **13**(2): 240-7.

Massey JT, Moore TF, Parsons VL, Tadros W (1989), Design and Estimation for the National Health Interview Survey, 1985 94, Vital Health Statistics 2(110).

Pamuk, E, Makuc D, et al. (1998), Socioeconomic Status and Health Chartbook. *Health United States*, 1998, NCHS Statistics.

R Development Core Team (2009), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

U.S. Decennial Life Tables for 1999-2001, United States Life Tables. NVSR Volume 57, Number 1. 37 pp. (PHS) 2008-1120, URL http://www.cdc.gov/nchs/products/life_tables.htm#decennial

Figure 1.

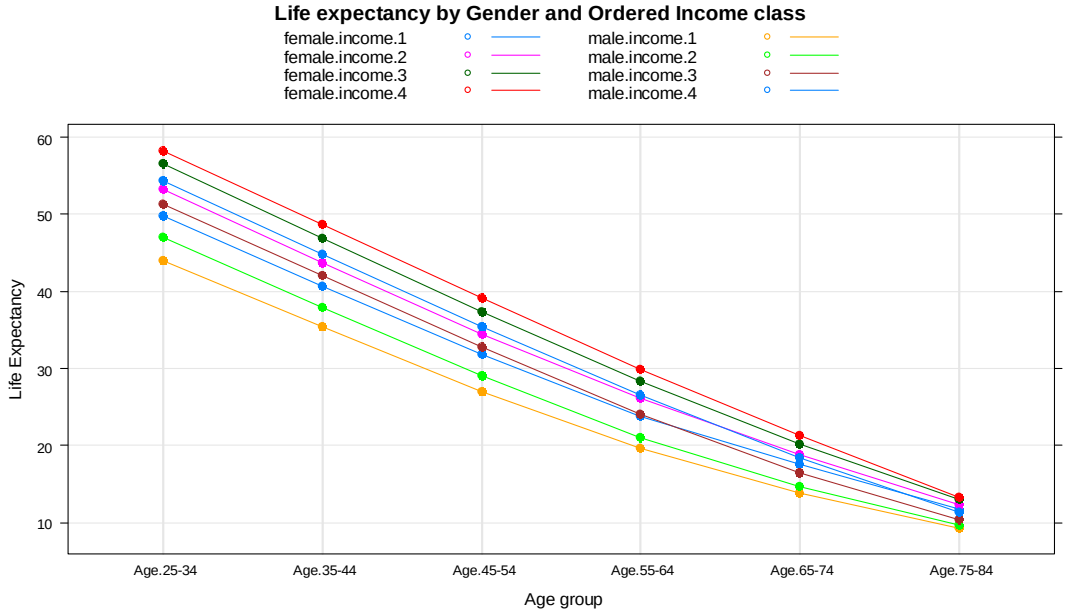

Figure 2.