HNUE JOURNAL OF SCIENCE DOI: 10.18173/2354-1059.2023-0055
Natural Science, 2023, Volume 68, Issue 3, pp. 13-21
This paper is available online at http://stdb.hnue.edu.vn

CONFIDENCE INTERVAL OF LIFE EXPECTANCY ESTIMATE

Nguyen Thanh Nga[1] and Ho Dang Phuc[2]
[1]*Faculty of Mathematics, Banking Academy of Viet Nam*
[2]*Institute of Mathematics, Vietnam Academy of Science and Technology*

Abstract. Concerning the local parametric estimation method, a novel method of life expectancy estimation, the article uses the bootstrap method to evaluate this method in extensive experiments with real-world datasets. The obtained results show that the local parameter method provides the narrower confidence intervals in comparison to the Chiang method. This provides the statistical tests to get higher performance in the detection of life expectancy estimate differences.
Keywords: life expectancy estimation methods, confidence interval, survival analysis, Weibull distribution.

1. Introduction

Life expectancy (LE) is the most remarkable summary index of the mortality experience of populations and is a key demographic indicator used to assess the population health. This indicator is used to make international comparisons between countries health [1]. This is also a good measure to examine geographic and socio-demographic inequalities in mortality [2]. The confidence intervals (CI's) of LE estimates can be compared to detect statistically significant differences.

The LE estimate method proposed by Chiang [3-4] is widely applicable (e.g., [2, 5, 6]). Using current life table creation based on survival theory, the Chiang method provides abridged life tables that aggregate deaths and population data into age groups under 1, 1-4, 5-9,... 80-84, 85 and over. Then Chiang provided a detailed assessment of the statistical properties of the LE estimate [4]. He derived the analytic equation for the variance of LE and its CI.

However, as it has been shown in the works [7-9], there are biases in Chiang's current life tables related to the estimates of age specific mortality rates and the conditional probabilities of dying in age intervals. Moreover, the ratio of the death amount and population number in 5-year age intervals usually provides a biased estimate

Received October 8, 2023. Revised: October 20, 2023. Accepted October 27, 2023.
Contact Nguyen Thanh Nga, e-mail address: ngant@hvnh.edu.vn

of age-specific mortality rates [7, 10-11]. These impact the accuracy of the Chiang LE estimate and yield a larger variance of the estimate, which makes related hypothesis tests to be less powerful.

The new method, called the "Local parametric method" (LP method), is based on the theoretical background of the survival process with local parametric Weibull distributions. The recent work [12] showed that this method can be applied to abridged datasets, showing certain advantages over the Chiang method. To further develop the above results, the main objective of this article is to use the bootstrap method to compare the life expectancy estimation results and provide confidence intervals for the Chiang method and LP method.

The rest of the article is organized as follows. Section 2 shows the theoretical analysis of the LP method. In Section 3, we experiment with the CI Bootstrap calculations for LE estimates on a longitudinal survival dataset of FilaBavi [13] using the Chiang method and the LP method. The conclusion is drawn in Section 4 with some discussions on the obtained results.

2. Life expectancy estimate by local parameter method

In [12], the LP method is based on the Weibull distribution with local parametrization in 19 age bands of $[0;1)$, $[1;5)$, $[5;10)$, ..., $[80;85)$, and $[85;\infty)$. For $j = 1,2,...,19$ let $[x_j, x_j + o_j)$ denote the j-th age band with length o_j. We set $o_1 = 1$; $o_2 = 4$; $o_j = 5$ for all $j = 3,4,...,18$; and $o_{19} = \infty$. Typically, data in population studies are comprised of follow-up reports over a certain calendar year, from January 1 to December 31 of the year. This article study, instead, uses abridged survival data that is somewhat different from conventionally reported abridged data. Specifically, instead of the midyear day (July 1), the last day (December 31) of the current observation year is used to determine the age of any person in the studied population. Let Y be a random variable that denotes the individual age counted on the last day (December 31) of the current observation year. Then Y is used to classify the age groups $[0;1)$, $[1;5)$, $[5;10)$, ..., $[80;85)$, and $[85;\infty)$ in abridged data. Then, except for the final age band, which is an open ended interval, all the early age bands are of the form $[x_j; x_j + o_j)$, for all $j = 1,2,...,19$. Let's recall that the random variable T indicates the time from birth to death of an individual. Then we can split T into the sum of random variables, each is defined on one of the age intervals $[x_j; x_j + o_j)$, $j = 1,2,...,19$, by

$$T = T \cdot \mathbf{1}_{[x_1;x_2)}(Y) + T \cdot \mathbf{1}_{[x_2;x_3)}(Y) + \ldots + T \cdot \mathbf{1}_{[x_{18};x_{19})}(Y) + T \cdot \mathbf{1}_{[x_{19};\infty)}(Y). \quad (2.1)$$

where $\mathbf{1}_A(x)$ the indicator function of set A.
For $j = 1,2,\cdots,19$, we model the random variable $T \cdot \mathbf{1}_{[x_j;x_j+o_j)}(Y)$ to have local Weibull distribution by formulas

$$T \cdot \mathbf{1}_{[x_1;x_2)}(Y) = T_1 \cdot \mathbf{1}_{[x_2;x_3)}(Y), T \cdot \mathbf{1}_{[x_j;x_j+o_j)}(Y) = (x_j - 1 + T_j) \cdot \mathbf{1}_{[x_j;x_j+o_j)}(Y). \quad (2.2)$$

for $j = 1, 2, \cdots, 19$, with the random variable T_j is supposed to have the Weibull density function

$$f_j(t) = k_j \lambda_j (\lambda_j t)^{k_j - 1} e^{-(\lambda_j t)^{k_j}} \quad for \;\; 0 \le t < \infty \, , \, \lambda_j > 0 \, . \tag{2.3}$$

The argument of [12] provides the estimation of the local scale parameters $\hat{\lambda}_j$, $j = 1, 2, ..., 19$ as follows:

$$\hat{\lambda}_1 = \left\{ \frac{-1 + \sqrt{1 + 4 \cdot \frac{k_1}{2k_1+1} \cdot \frac{death_1}{n_1}}}{\frac{2k_1}{2k_1+1}} \right\}^{1/k_1} . \tag{4}$$

$$\hat{\lambda}_2 \approx \left\{ \frac{4(k_2+1)(death_2 + d_{1;2})}{n_2 \cdot [(k_2+1)5^{k_2} - 1]} \right\}^{1/k_2} , \tag{5}$$

where

$$d_{1;2} \approx n_1 \times \big(1 - \frac{\hat{\lambda}_1^{k_1}}{k_1 + 1} + \frac{\hat{\lambda}_1^{2k_1}}{2(2k_1+1)} - e^{-\hat{\lambda}_1^{k_1}} \big) \, .$$

For the scale parameter λ_j, $j = 3, 4, ..., 18$,

$$\hat{\lambda}_j \approx \left\{ \frac{o_j(k_j+1)(death_j + d_{j-1;2})}{n_j \cdot [(k_j+1)(o_j+1)^{k_j} - 1]} \right\}^{1/k_j} , \tag{6}$$

with

$$d_{2;2} \approx \frac{n_2}{4} \times \Big(\frac{-5^{k_2+1} + 4^{k_2+1}}{k_2+1} + 5^{k_2} \Big) \hat{\lambda}_2^{k_1}$$

and

$$d_{j;2} \approx \frac{n_j}{o_j} \times \Big(\frac{-(o_j+1)^{k_j+1} + o_j^{k_j+1}}{k_j+1} + (o_j+1)^{k_j} \Big) \lambda_j^{k_j} \, .$$

For the open-ended age group $[85; +\infty)$, with $d_{19} = death_{19} + d_{18;2}$, the estimation of the scale parameter λ_{19} is given as

$$\hat{\lambda}_{19} = -\ln\Big(1 - \frac{d_{19}}{n_{19}}\Big) = \ln\Big(\frac{n_{19}}{n_{19} - d_{19}}\Big) \, . \tag{7}$$

As mentioned in [12], the shape parameter k_j of the random variable W_j influences the mortality rate in $[x_j, x_j + o_j)$, $j = 1, 2, \ldots, 19$. The mortality rate in this model, which can be interpreted as $k_j \lambda_j (\lambda_j t)^{(k_j - 1)}$, decreases over time in the first year of life, but stays constant with time for the medium age groups. This mortality rate increases over time in the last age groups. In the LP method, the value of k_j must be pre-determined and used consistently for any dataset. Specifically, following our suggestion in [12], we propose the set of parameters for k_j as follows:

$$\{k_i\} = \{0.1; 0.2; 0.9; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 2; 2; 2; 1\}$$

where $k_1 = 0.1; k_2 = 0.2; k_3 = 0.9; k_j = 1$ for the medium age groups with $j = 4, \ldots, 15; k_j = 2$ for the three older age groups with $j = 16, 17, 18$; and $k_{19} = 1$.

Using the parameters $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_{19}$ estimated by (4)-(7) and sequence of shape parameters'values k_j, the argument of [12] provides a method to get life expectancy estimation.

In 1979, Brad Efron [14] invented a revolutionary new statistical procedure called the bootstrap method. This is a computer-intensive procedure that substitutes fast computation for theoretical math. The main benefit of the bootstrap is that it allows statisticians to set confidence intervals on parameters without having to make unreasonable assumptions. Specifically, it creates n resamples (with replacement) from a single set of observations, and calculates the statistic of interest for each of these resamples. Then, bootstrap percentile CI of $\hat{\theta}$ (an estimator of θ) can be obtained as follows:

- n random bootstrap samples are generated,
- a parameter estimate is calculated from each bootstrap sample,
- all n bootstrap parameter estimates are ordered from the lowest to highest,
- and the CI is constructed as follows

  $$[\hat{\theta}_{lowerlimit}, \hat{\theta}_{upperlimit}] = [\hat{\theta}_j, \hat{\theta}_k]$$

  where $\hat{\theta}_j$ denotes the j-th quantile (lower limit), and $\hat{\theta}_k$ denotes the k-th quantile (upper limit). For example, a 95% percentile bootstrap CI with 1,000 bootstrap samples is the interval between the 25-th quantile value and the 975-th quantile value of the 1000 bootstrap parameter estimates.

To evaluate the accuracy of life expectancy between the LP method and the Chiang method, in this article, we use the bootstrap method to calculate the CI for life expectancy estimation. The results are shown in the following.

3. Using bootstrap method for evaluating life expectancy estimation

In this section, a real survival dataset of FilaBavi [13] is used to compare the effectiveness of the LP method for life expectancy estimate and the traditional Chiang method. The 95% boostrap confidence intervals of the Chiang life expectancy estimate and the LP life expectancy estimate are computed to compare the performance of the two estimate methods.

Table 1. Population and death numbers in annual abridged data sets

Year	F-Pop	F-Death	M-Pop	M-Death	Total Pop	Total Death
2000	25821	115	23974	126	49795	241
2001	25680	124	23888	134	49568	258
2002	25610	108	23734	147	49344	255
2003	24428	111	23584	130	49012	241
2004	25410	118	23673	149	49083	267
2005	25455	130	23691	166	49146	296
2006	25535	131	23895	154	49430	285
2007	25688	123	24025	148	49713	271
2008	25848	143	24114	169	49962	312
2009	25927	140	24264	174	50191	314
2010	26185	145	24638	177	50824	322
2011	26198	162	24772	180	50970	342
2012	26315	149	25029	196	51344	345
2013	26327	132	25155	163	51482	295
2014	26776	142	25545	172	52323	314

The dataset of FilaBavi includes the first interviews in 1999, March, and the last interviews in 2015, October. That means there are 15 years (2000 to 2014) of complete observation recorded in this dataset. Extracting from the dataset, this study uses the data file that has been reorganized by splitting into 15 one-year observation semi-cohort data files, each of them related to one specific year among the 2000 to 2014 years. Each dataset of the 15 one-year data files is used to produce the respective abridged data file. For the abridged dataset of each year, we draw 500 bootstrap samples of size 20000, these samples are used to calculate life expectancy estimates of the Chiang method and of the LP method. Consequently, we get 500 values of life expectancy estimation for each year, then we obtain 95% CI of the Chiang life expectancy estimate and of the LP life expectancy estimate. A brief description of these abridged datasets is given in Table 1. The ”F-Pop” and ”F-Death” columns in this table represent the pairs of female population and female death numbers in each year, respectively. The columns ”M-Pop” and ”M-Death” contain the pairs of male population and male death numbers, whilst the columns ”Total Pop” and ”Total Death” show the total numbers of annual population and deaths.

To check the distribution of life expectancy local parametric estimation. We draw the histogram of these estimations for each year (e.g., Figure.1). These results showed that the distribution of life expectancy estimations by the LP method has a normal distribution.

Tables 2 and 3 represent the sample mean of 500 values of life expectancy estimation (columns ”LE”) together with their 95% bootsrap confidence intervals (columns ”95% CI of Ch”, ”95% CI of LP”) calculated from the above mentioned

Table 2. Bootstrap confidence interval by the Chiang method and the LP method for male

Year	Mean of Ch. Est	95% CI of Ch	Mean of LP	95% CI of LP
2000	73.19	70.91;75.47	72.89	71.35;74.43
2001	72.11	69.81;74.40	71.82	70.31;73.32
2002	71.68	69.53;73.84	71.57	70.22;72.92
2003	73.09	70.80;75.38	72.76	71.30;74.22
2004	71.87	69.54;74.20	71.47	69.91;73.03
2005	70.62	68.24;73.01	70.22	68.63;71.82
2006	72.11	69.81;74.42	71.85	70.37;73.34
2007	73.22	70.81;75.62	72.74	71.09;74.39
2008	71.63	69.40;73.85	71.32	69.90;72.75
2009	71.57	69.38;73.76	71.23	69.81;72.65
2010	71.96	69.67;74.24	71.52	69.98;73.07
2011	72.15	69.94;74.35	71.74	70.26;73.23
2012	71.51	69.43;73.59	71.35	70.07;72.64
2013	74.27	72.08;76.46	73.88	72.44;75.32
2014	74.60	72.70;76.51	74.48	73.29;75.67

FilaBavi dataset, applying the Chiang method and the LP method for males and females.

Table 3. Bootstrap confidence interval by the Chiang method and the LP method for female

Year	Mean of Ch. Est	95% CI of Ch	Mean of LP	95% CI of LP
2000	81.03	78.80;83.27	80.27	78.90;81.65
2001	81.19	79.16;83.23	80.52	79.29;81.74
2002	82.91	80.56;85.27	82.01	80.54;83.47
2003	82.59	80.51;84.66	81.71	80.48;82.95
2004	83.33	81.29;85.38	82.52	81.23;83.82
2005	82.84	80.89;84.78	82.11	80.95;83.27
2006	83.17	80.90;85.45	82.29	80.88;83.70
2007	84.67	82.68;86.65	83.93	82.80;85.06
2008	83.27	81.29;85.24	82.78	81.66;83.91
2009	83.78	81.89;85.66	83.20	82.14;84.26
2010	83.93	82.01;85.85	83.36	82.25;84.47
2011	82.77	80.98;84.55	82.18	81.18;83.19
2012	84.23	82.53;85.94	83.62	82.66;84.58
2013	85.31	83.37;87.25	84.54	83.40;85.68
2014	86.21	84.62;87.80	85.56	84.62;86.50

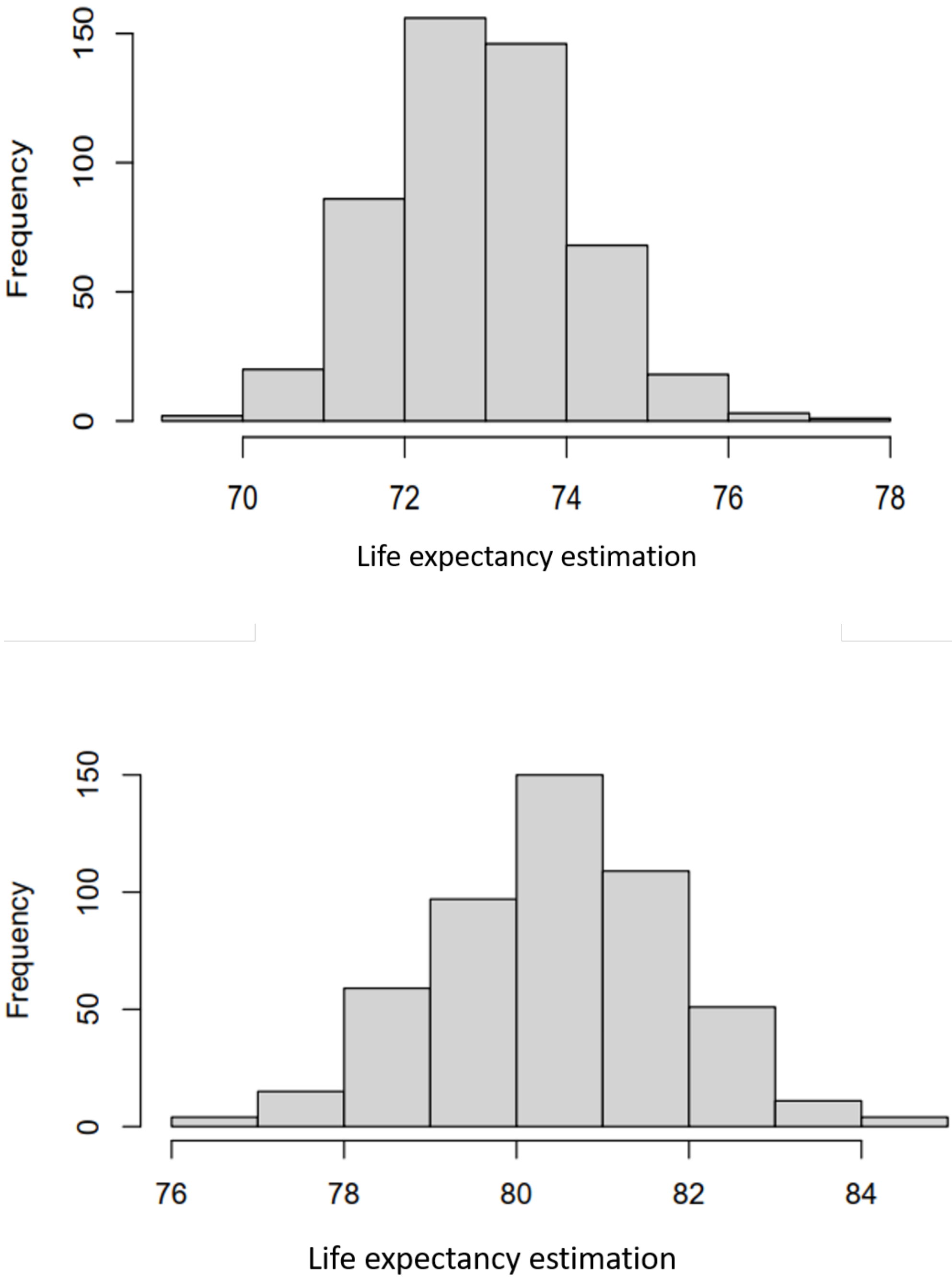

Figure 1. Histogram of life expectancy estimation by LP method for male (top) and female (bottom) in Year 2000

Comparing the figures in Table 2, Table 3 we observe that in most cases, the Chiang method gives over-estimated than the LP method. This is appropriate with the results shown in [12]. Additionally, the LP method always provides the narrower confidence intervals of LE estimates than the Chiang method does. In detail, for the male population, the width of LP confidence intervals is equal to around 70% of that derived from the Chiang method. Simultaneously, for the female population, the ratios of the width of LP confidence intervals to the corresponding width yielded from the Chiang method are equal to about 65%. The comparison results point out the advantages of the LP life expectancy estimate method over the Chiang estimate method when showing its higher effectiveness in the statistical tests to detect the differences of life expectancy estimates.

4. Conclusions

Life expectancy as a summary index of population health is widely used by epidemiologists, ecologists, economists, and policy makers, to examine the geographic and socio-demographic inequalities in health and to make international comparisons between countries' living standards. The statistical tests of LE estimates' comparisons are usually based on the estimates' confidence intervals. Up to now, the Chiang method has been the most popularly used tool to get LE estimates. However, many works like [7, 9-11] have shown that there are biases in the Chiang life expectancy estimate. In more detail, Scherbov and Ediev [15] have proved that the Chiang method provides an overestimate of life expectancy. The bias makes the estimate variance to be larger, which makes the attached confidence intervals to be wider.

The experiments presented in Section 3 based on real data demonstrate the advantages of this LE estimate method over the Chiang method. Especially, the narrower confidence intervals of the new method estimate presented in the calculation results confirm the higher precision level of the respective LE estimates. Simultaneously, these narrower confidence intervals allow statistical tests to have a higher power in the detection of significant differences in LE estimate. These facts also expose the higher effectiveness of the LE estimate done by the LP method.

However, some theoretical details should be clarified to strengthen the advantages of the local parametric estimation method. An open problem is to determine the formula for the variance of the life expectancy estimated by the local parametric estimation method. This problem should be put for further investigation. Although the need for improvement, the proposed local parametric method of life expectancy estimation can be applied instead of the ordinary method of Chiang, because of the mentioned advantages of the new method.

Acknowledgements. The analyses in Section 3 have been conducted by using the longitudinal dataset of FilaBavi - Ba Vi Epidemiological Field Laboratory. Thanks are due to Prof. Dr. Nguyen Thi Kim Chuc, the main coordinator of FilaBavi, and to other scientists and workers who participated in conducting many years of surveys and processing to produce the dataset, which plays a critical role in this study.

REFERENCES

[1] J.M. Eyler, 1979. Victorian social medicine. The ideas and methods of William Farr, *Baltimore: Johns Hopkins University Press*.

[2] C. Griffiths, J. Fitzpatrick, 2001. Geographic inequalities in life expectancy in the United Kingdom, 1995-1997, *Health Statistics Quarterly*, 09, pp. 16-28.

[3] C.L. Chiang, 1972. On Constructing Current Life Tables, *Journal of the American Statistical Association*, 67:339, pp. 538-541.

[4] C.L. Chiang, 1984.*The Life Table and its Applications*, Malabar (FL), Robert E Krieger Publ Co.

[5] J.P. Cristia, 2009. Rising mortality and life expectancy differentials by lifetime earnings in the United States, Working Paper No. 665, *Inter-American Development Bank, Research Department, Washington, DC*.

[6] R.M. Kaplan, M.L. Spittel, T.L. Zeno, 2014. Educational Attainment and Life Expectancy, *Policy Insights from the Behavioral and Brain Sciences*, Vol.1, pp. 189-194.

[7] J.H. Pollard, 1989. On the derivation of a full life table from mortality data recorded in five-year age groups, *Mathematical Population Studies: An International Journal of Mathematical Demography*, 2:1, pp. 1-14.

[8] J.R. Wilmoth, K. Andreev, D. Jdanov, D.A. Glei, 2007. Methods Protocol for the Human Mortality Database, *University of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock*.

[9] J.J. Hsieh, 1991. A general theory of life table construction and a precise abridged life table method, *Biometrical Journal*,33: (2), pp. 143-162.

[10] A.L. Golbeck, 1986. Probabilistic Approaches to Current Life Table Estimation, *The American Statistician*, 40:3, pp. 185-190.

[11] A. Meulen, 2012. Life tables and Survival analysis, *Statistics Netherlands*, The Hague/ Heerlen.

[12] P. Ho Dang, T.N. Nguyen, 2022. New methods of life expectancy estimation, *Environmental and Ecological Statistics*,29, pp. 587-606.

[13] T.K.C. Nguyen, K.D. Vinod, 2003. FilaBavi, a demographic surveillance site, an epidemiological field laboratory in Vietnam, *Scand J. Public Health*, 31, (Suppl. 62), pp. 3-7.

[14] B. Efron, R. J. Tibshirani, 1994. *An introduction to the bootstrap*, CRC press.

[15] S. Scherbov, D. Ediev, 2012. Significance of life table estimates for small populations: Simulation-based study of standard errors, *Demographic Research*, 24:22, pp. 527-550.