A PRECISE LIFE TABLE METHOD

John J. Hsieh, University of Toronto

1. INTRODUCTION

This paper provides a precise method of constructing abridged life tables. Such construction involves two problems: The main one is the estimation of the survival rate, ${}_np_x = l(x+n)/l(x)$, from deaths registered during a given base period and populations enumerated or estimated at mid-years in each age interval; the secondary problem is the estimation of the stationary population ${}_nL_x$.

The estimation of the survival rate calls for the solution of certain equations which relate the observed age-specific death rate to the function underlying the age distribution in the stationary population on the one hand and the lifetime distribution in the stationary population on the other (Section 2). The solution of these equations could be regarded as an approximation of a dimensionless function by known dimensional functions. Keyfitz and Frauenthal (1975) solved such an equation and obtained an explicit functional relationship which approximated the survival rate in terms of age-specific death rates and mid-year populations, and which they showed are considerably more accurate than those derived by using the age distribution of the stationary population (Greville 1943). The main purpose of this paper is to provide a different set of explicit formulas (Section 3) which will be shown to be more accurate than Keyfitz-Frauenthal's and capable of removing the two defects inherent in their method (Section 4). A complete cubic spline obtained from consideration of the lifetime distribution is used to compute the ${}_nL_x$ function and the result is shown to be more accurate than other existent methods (Section 5). Life tables so constructed are to be viewed as constructed at the midpoint of the base period. A detailed discussion of application of spline functions to life table construction, including the construction of complete life tables, is given in the more comprehensive paper, Hsieh (1977).

2. FUNDAMENTAL CONCEPTS AND EQUATIONS

We shall use a star superscript (*) to distinguish functions in the observed population from their corresponding functions in the stationary population. Let $l^*(x,t)$ be continuous having a continuous first order partial derivative with respect to age x, and represent the profile of the observed population pyramid at calendar time t (the unit of $l^*(x,t)$ is "persons per year") so that $l^*(x,t)dx$ is the number of individuals aged x to x+dx at time t and $l^*(x,t)$ dxdt is the number of individual-time units observed on the region dxdt. Let $\mu^*(x,t)$, which possesses similar regularity properties to $l^*(x,t)$ be the force of mortality at age x and calendar time t ($\mu^*(x,t)$ has unit "per year"). Then, the death rate ${}_n^hM_x^t$ (whose unit is "per year") for the age interval [x,x+n) and the base period [t,t+h] (usually n=5 or 4 years, and h=3 or 1 year.) can be expressed as

$$ {}_n^hM_x^t = \frac{\int_0^h\int_0^n l^*(x+v,t+u)\ \mu^*(x+v,t+u)\,dvdu}{\int_0^h\int_0^n l^*(x+v,t+u)\,dvdu} . \tag{2.1}$$

The numerator in the above expression (whose unit is "persons") represents the number of individuals aged x to x+n who die during the base period [t,t+h] and is known from the death data. The denominator (whose unit is "person years") represents the individual-time units of exposure to the risk of death in the same age interval and base period and is unknown because the inner integral, which is the population between the ages x and x+n at any time point in the base period, is unknown except at midyears.

By definition, to construct a life table at the mid-period is to take the hazard function $\mu^*(x,t+h/2)$ of the observed population at this time point to be the hazard function $\mu(x)$ for the stationary population of the life table. Consider the time variable to be fixed at the midpoint t+h/2 of the base period and write

$$ {}_nP_x \equiv \int_x^{x+n} l^*(v,t+h/2)\,dv. $$

Then, (2.1), with the mid-period time point t+h/2 understood, can be written as

$$ \int_x^{x+n} l^*(v)\mu(v)\,dv = {}_nM_x\,{}_nP_x . \tag{2.2}$$

The person-year integral in the denominator of (2.1) can be numerically integrated and expressed in terms of populations at mid-years (Hsieh, 1976). ${}_n^hM_x^t$ can then be calculated from the observed data.

From the lifetime distribution theory or the pure death process we have for the lifelength X at midperiod,

$$ {}_np_x \equiv \Pr\{X>x+n \mid X\geq x\} = \exp\{-\int_x^{x+n} \mu(v)\,dv\}. \tag{2.3}$$

(2.3) expresses the survival rate ${}_np_x$ in terms of the force of mortality $\mu(x)$ over the corresponding age interval in the stationary population.

3. CALCULATION OF THE SURVIVAL RATE

In (2.2) both quantities on the right hand side are known whereas both $l^*(v)$ and $\mu(v)$ on the left hand side are unknown functions. This equation can be regarded as an integral equation with $\mu(v)$ as the unknown function. Once $\mu(v)$ is solved for, ${}_np_x$ is obtained from (2.3). Equation

(2.2) can be shown to be indeterminate (Hsieh, 1977). Thus, in order to uniquely determine $\mu(x)$ or the integral

$$\int_x^{x+n} \mu(v)\,dv$$

from (2.2), it is necessary to impose constraints either on $l^*(x)$ or on $\mu(x)$ alone or on both $l^*(x)$ and $\mu(x)$.

Aside from the standard regularity properties imposed on $l^*(x)$ in Section 2 for mathematical convenience, $l^*(x)$ also has its natural demographic properties: $l^*(x) \geq 0$ for $0 \leq x < \omega$, and $l^*(\omega) = 0$, where ω is the maximum age. Now, letting

$$H(x) \equiv \int_x^{\omega} l^*(v)\,dv$$

and integrating by parts on the left hand side of (2.2) yields

$$\int_x^{x+n} l^*(v)\mu(v)\,dv = \mu(x)H(x) - \mu(x+n)H(x+n) + \int_x^{x+n} \mu'(v)H(v)\,dv, \tag{3.1}$$

where the prime signifies derivative. A Taylor expansion on $H(v)$ about $v = x+n/2$ in linear terms gives

$$H(v) = H(x+n/2) - l^*(x+n/2)(v-x-n/2) + E(v), \tag{3.2}$$

with error term

$$E(v) = -\int_{x+n/2}^{v} (v-y)\, l^{*\prime}(y)\,dy.$$

Entering (3.2) into the last term on the right hand side of (3.1) and carrying out the integration, we have

$$\int_x^{x+n} \mu'(v)H(v)\,dv = H(x+n/2)[\mu(x+n)-\mu(x)] - (n/2)\,l^*(x+n/2)[\mu(x+n)-\mu(x)] + l^*(x+n/2)\int_x^{x+n}\mu(v)\,dv + \int_x^{x+n}\mu'(v)E(v)\,dv. \tag{3.3}$$

We shall now approximate the integral involving the error term $E(v)$. Using the integral expression for the error term in (3.2) in the last integral of (3.3), replacing one of the two functions which form the product in the integrands by its average value, and reversing the order of integration in the iterated integral gives:

$$\int_x^{x+n} E(v)\mu'(v)\,dv = -\int_x^{x+n}\left\{\int_{x+n/2}^{v}(v-y)\,l^{*\prime}(y)\,dy\right\}\mu'(v)\,dv$$

$$\simeq -(1/n)\left\{\int_x^{x+n}\int_{x+n/2}^{v}(v-y)\,l^{*\prime}(y)\,dy\,dv\right\} \times \left\{\int_x^{x+n}\mu'(v)\,dv\right\}$$

$$= -(1/n)\left[\int_{x+n/2}^{x+n}\left\{\int_y^{x+n}(v-y)\,dv\right\}l^{*\prime}(y)\,dy + \int_{x+n/2}^{x}\left\{\int_x^{y}(v-y)\,dv\right\}l^{*\prime}(y)\,dy\right][\mu(x+n)-\mu(x)]$$

$$\simeq (n/24)[l^*(x)-l^*(x+n)][\mu(x+n)-\mu(x)]. \tag{3.4}$$

Combining (2.2), (2.3), (3.1), (3.3) and (3.4), and using $H(x)-H(x+n) \equiv {}_nP_x$, yields

$$\ln {}_np_x = -[1/l^*(x+n/2)]\Big[{}_nM_x\,{}_nP_x - \mu(x)\,{}_nP_x + (n/2)\,l^*(x+n/2)\{\mu(x+n)+\mu(x)\} - [H(x+n/2)-H(x+n) + (n/24)\{l^*(x)-l^*(x+n)\}] \times \{\mu(x+n)-\mu(x)\}\Big]. \tag{3.5}$$

The approximations in (3.4) made use of the mathematical fact that the two functions

$$G_1(y) = \int_y^{x+n}(v-y)\,dv \quad\text{and}\quad G_2(y) = \int_x^{y}(v-y)\,dv$$

do not change sign in any age interval $[x, x+n)$, and the demographic fact that $\mu'(v)$ does not change sign except for one or three intervals where the relative minima or maximum of the $\mu(v)$ curve occur. In these intervals, however, the value of

$$\int_x^{x+n}\mu'(v)\,dv = \mu(x+n)-\mu(x)$$

is near zero and therefore the approximation has little effect on the result. At worst, the approximation may be regarded as taking the error $E(v)$ to be constant within each of these transition intervals.

Our next task is to approximate by numerical methods the unknown quantities that appear in (3.5) in terms of mid-period populations and death rates. We adopt the conventional division of the whole agespan for the abridged life table into 0, 1, 5, 10, ..., 85, 90, ω years, where ω is the maximum age to which any individual can live. The prsent proposed method may be used to advantage for wider age groups. However, data for single-year age intervals, even when available, are not reliable; if they were, many simple methods would produce life table functions about as accurate as those produced by sophisticated methods such as the present one.

Our life table method begins with age one and ends at the exact age marking the start of the terminal age interval (90 in this case). The precise method for the first year of life, because of gross underenumeration (and estimation) of infants, requires birth data and is therefore different from the method for ages beyond one (see Greville 1947). Life table functions for the terminal age interval, because of the unknown ω, are conventionally computed using the fact that $l(\omega)=0$ and the assumption that the age distribution of the observed population is identical with that of the stationary population.

The formula for computing n-year survival rate is as follows:

$$\ln {}_np_x = -n\,{}_nM_x - n\,A_xB_x\,/\,{}_nP_x\,, \tag{3.6}$$

where (i) for $x = 1$, n is 4 and

$$A_1 = (725\,{}_4P_1 - 418\,{}_5P_5 - 162\,{}_5P_{10})/12825,$$

$$B_1 = (475\,{}_4M_1 + 722\,{}_5M_5 - 114\,{}_5M_{10})/1083 - (365/31)D_m/(B-D_1+D_m), \quad \text{or}$$

$$B_1 = (-1120\,{}_4M_1 + 1444\,{}_5M_5 - 324\,{}_5M_{10})/855;$$

(ii) for $x = 5, 10, \ldots, 75$, n is 5 and

$$A_x = (9\,{}_5P_{x-5} - 3\,{}_5P_x - 5\,{}_5P_{x+5} - {}_5P_{x+10})/192,$$

$$B_x = (-3\,{}_5M_{x-5} - 3\,{}_5M_x + 7\,{}_5M_{x+5} - {}_5M_{x+10})/8;$$

and (iii) for $x = 80, 85$, n is 5 and

$$A_x = ({}_5P_{x-10} + 2\,{}_5P_{x-5} - 3\,{}_5P_x)/48,$$

$$B_x = ({}_5M_{x-10} - 4\,{}_5M_{x-5} + 3\,{}_5M_x)/2.$$

Formula (3.6) is obtained by using collocation polynomials, ${}_nP_x \equiv H(x)-H(x+n)$ and the approximations ${}_nP_x = nl^*(x+n/2)$ and ${}_nM_x = \mu(x+n/2)$ in equation (3.5). For age intervals other than the first, the following general form of Newton's formulas with various chosen values of j, r and s,

$$f_{x+(j+r)n} = \sum_{i=0}^{s} \frac{r(r-1)\ldots(r-i+1)}{i!} \Delta^i f_{x+jn} + E_t,$$

where $\Delta^i f_x$ designates the ith forward difference of f_x and E_t denotes the truncation error, were used to express the unknown functions in (3.5) as linear combinations of mid-period populations and death rates. Because unequal age intervals were involved, Lagrange's formulas for collocation polynomials were employed for the first age interval ($x=1$, $n=4$). Also, the abrupt bend of the $\mu(x)$ curve around age one renders it inappropiate, except for countries with very low infant mortality, to extrapolate $\mu(1)$ in terms of death rates in succeeding age intervals. Since $l(x)$ is convex at age one, $\mu(1)$ is closely estimated by the ratio of the conditional probability of dying in the 12th month of life to the length of the month:

$$\mu(1) = (365/31)D_m/[B - D_1 + D_m], \qquad (3.7)$$

where D_m, D_1 and B, respectively, denote the number of deaths in the 12th month of life, the number of deaths under one year, and the number of births, all during the base period. The data for D_m and D_1 are given for various countries in the 1974 U.N. Demographic Yearbook. Utilization of (3.7) leads to the alternative expression for B_1 given in (3.6).

4. COMPARISON OF ACCURACY

Keyfitz and Frauenthal (1975) showed that their life table method is more accurate than other ones. In this section we emphasize comparisons between the new method and the Keyfitz-Frauenthal (denoted henceforth as "K-F") method.

To effect a precise comparison of accuracy, we use the test proposed by Keyfitz and Frauenthal (1975) which assumes both functions $l^*(x)$ and $l(x)$ to be known, where $l(x)$ is the number of survivors to age x out of $l(0)$ births in the life table so that $l(x+n)=l(x)\,{}_np_x$. Adopting K-F stable population profile $l^*(x)$ and Makeham's graduation formula for $l(x)$,

$$l^*(x) = 10^6[1-\exp(x/100-1)] \qquad (4.1)$$

$$\text{and} \quad \ln l(x) = \ln l(0) + x \ln s + (c^x-1) \ln g, \qquad (4.2)$$

where $s=.999859$, $g=.999743$ and $c=1.109887$; $\mu(x)$, $\mu'(x)$, $l^{*\prime}(x)$, ${}_nM_x$ and ${}_nP_x$ are computed using (2.3) and (2.2).

Next, the new formula (3.6) and the formulas for the following abridged life table methods:

Greville:
(1943)

$$\ln {}_np_x = -n\,{}_nM_x - n^2\,{}_nM_x({}_nM_{x+n} - {}_nM_{x-n})/24, \qquad (4.4)$$

Reed and Merrell:
(1939)

$$\ln {}_np_x = -n\,{}_nM_x - .008n^3\,{}_nM_x^2, \qquad (4.5)$$

Keyfitz and Frauenthal:
(1975)

$$\ln {}_np_x = -n\,{}_nM_x + n({}_nP_{x+n} - {}_nP_{x-n})({}_nM_{x+n} - {}_nM_{x-n})/(48\,{}_nP_x) \qquad (4.6)$$

are applied to the synthetic ${}_nM_x$ and ${}_nP_x$ to reproduce the life table $l(x)$. Since K-F formula (4.6) cannot be used for computing ${}_np_x$ for the initial interval [0,5), the simple formula $\ln {}_np_x = -n\,{}_nM_x$ obtained from (2.2) and (2.3) by assuming constant force of mortality within this age interval, is used to compute $l(5)$ for all life table methods. The results are shown in Table 1. The cumulative absolute errors are found to be <u>4.55</u> for the new formula (3.6), <u>41.71</u> for the K-F formula (4.6), <u>825.66</u> for the Reed and Merrell formula (4.5) and <u>996.18</u> for the Greville formula (4.4).

The principal advantage of the new method over the K-F method is that the latter requires estimation of $l^{*\prime}(x)$ and $\mu'(x)$ while the former requires estimation of $l^*(x)$, $\mu(x)$ and

$$\int_{x+n/2}^{x+n} l^*(v)\,dv.$$

The well known fact that approximate derivatives

Table 1. Comparison of Exact Makeham $l(x)$ with Results of Four Approximate Life Table Methods

Age x	exact	Hsieh (3.6)	Keyfitz & Frauenthal (4.6)	Reed & Merrell (4.5)	Greville (4.4)
0	100000	100000	100000	100000	100000
5	99912	99912	99912	99912	99912
10	99812	99812	99812	99812	99812
15	99692	99692	99692	99692	99692
20	99538	99538	99538	99538	99538
25	99327	99327	99327	99328	99328
30	99021	99021	99021	99022	99022
35	98555	98555	98555	98556	98556
40	97822	97822	97821	97825	97825
45	96646	96646	96646	96652	96652
50	94744	94744	94743	94753	94753
55	91668	91668	91667	91684	91683
60	86754	86754	86752	86778	86776
65	79104	79104	79101	79134	79129
70	67747	67747	67741	67767	67754
75	52207	52208	52200	52176	52148
80	33679	33681	33670	33531	33481
85	16105	16107	16096	15828	15762
90	4651	4651	4647	4394	4346

Table 2. Comparison of Exact ${}_nL_x$ Computed from Makeham $l(x)$ with Results of Four Approximate Integration Methods

Age x	exact	(5.1)	Keyfitz & Frauenthal (8.2)	Polynomial (8.3)	Simple ratio (8.4)
1	399792	399790	399790	399790	399860
5	499316	499316	499316	499316	499520
10	498770	498770	498771	498771	499079
15	498092	498092	498093	498093	498532
20	497193	497193	497194	497194	497776
25	495920	495921	495923	495923	496648
30	494023	494024	494027	494027	494882
35	491082	491082	491089	491088	492049
40	486402	486403	486414	486413	487441
45	478855	478856	478873	478870	479909
50	466638	466640	466667	466656	467615
55	446992	446994	447038	447008	447748
60	415995	415996	416064	415988	416316
65	368844	368839	368942	368771	368461
70	301562	301546	301685	301379	300247
75	215361	215324	215488	215141	213169
80	122917	122916	123014	122987	120469
85	48574	48619	48613	49077	46870
Cumulative absolute error		114	677	1134	16347

Table 3. Abridged Life Table for Male Population: Canada, 1970-72

Age Group x (1)	${}_nP_x$ (2)	${}_nD_x$ (3)	${}_nM_x$ (4)	${}_nq_x$ (5)	$\ell(x)$ (6)	${}_nd_x$ (7)	${}_nL_x$ (8)	T(x) (9)	e(x) (10)
Under 1	182195	11173	0.020441	0.020022	100000	2002	98226	6933697	69.337
1-4	747410	2119	0.000945	0.003800	97998	372	391106	6835470	69.751
5-9	1152430	1913	0.000553	0.002843	97625	278	487398	6444365	66.011
10-14	1181450	1837	0.000518	0.002595	97348	253	486205	5956967	61.193
15-19	1074430	4697	0.001457	0.007292	97095	708	483891	5470762	56.344
20-24	941775	5266	0.001864	0.009267	96387	893	479666	4986871	51.738
25-29	800710	3556	0.001480	0.007369	95494	704	475669	4507205	47.199
30-34	660875	3287	0.001658	0.008271	94790	784	472058	4031536	42.531
35-39	645045	4243	0.002193	0.010911	94006	1026	467645	3559478	37.864
40-44	640765	6886	0.003582	0.017771	92981	1652	461080	3091833	33.252
45-49	613415	10406	0.005655	0.027980	91328	2555	450757	2630753	28.805
50-54	518895	14562	0.009354	0.045945	88773	4079	434378	2179996	24.557
55-59	472415	20730	0.014627	0.070894	84694	6004	409427	1745617	20.611
60-64	381690	26571	0.023205	0.110425	78690	8689	372915	1336191	16.980
65-69	296050	31482	0.035447	0.163899	70001	11473	322435	963276	13.761
70-74	205575	32751	0.053105	0.235759	58528	13798	258880	640840	10.949
75-79	139995	33145	0.078919	0.330026	44729	14762	186786	381961	8.539
80-84	85680	30650	0.119242	0.456339	29967	13675	114579	195175	6.513
85-89	40625	21181	0.173793	0.592992	16292	9661	55166	80595	4.947
90+	13940	10905	0.260760	1.000000	6631	6631	25430	25430	3.835

obtained from collocation polynomials conglomerate much larger errors than do approximations of functions and their integrals is reflected in the ample difference (41.71 versus 4.55) in the cumulative absolute error between the two life table methods based on the results of Table 1. With the transition from synthetic to real data, the K-F method would suffer still greater loss in accuracy than the new method. This is because the analytic curve used in the test may be close to the true curve and yet the two curves still may have very different slopes. Thus, for age distributions with dents and bulges such as those resulting from the two World Wars, the estimated values of $l^{*'}(x)$ in age intervals adjacent to where the dents and bulges occur, could differ vastly from the true values of $l^{*'}(x)$.

Another advantage of the present method over the K-F method regards coverage of the agespan. Since the K-F method requires three consecutive age intervals of equal length to calculate the survival rate for the central age interval, the K-F formula (4.6) cannot be used to compute survival rates for both the first age interval, either [0,5) or [1,5), and the last age interval [85,90). On the other hand, with no problem of estimation of slopes, the new formula (3.6) covers these two intervals just as well as other age intervals.

5. COMPUTATION OF STATIONARY POPULATION

With values of the survivorship function $l(x)$ available at the age points x=1, 5, 10, ..., 90, we now turn to the problem of computing stationary populations

$$ {}_nL_x = \int_0^n l(x+v)\,dv. $$

Various methods of approximating this integral exist in the literature with varying degrees of accuracy (see Table 2). We use the method of splines to approximate the integral

$$ L_i \equiv {}_{n_i}L_{x_i},\quad i=0,1,\ldots,k, $$

by the formula:

$$ L_i = n_{i+1}(l_i+l_{i+1})/2 + n_{i+1}^2(s_i-s_{i+1})/12, \quad (5.1) $$

where $l_i \equiv l(x_i)$ and the slopes $\{s_i\}$ are to be determined by solving the following system of k-1 equations for cubic splines:

$$ n_{i+1}s_{i-1} + 2(n_{i+1}+n_i)s_i + n_i s_{i+1} $$

$$ = 3\left[\frac{n_i}{n_{i+1}}(l_{i+1} - l_i) + \frac{n_{i+1}}{n_i}(l_i-l_{i-1})\right], \quad (5.2) $$

for i=1,2,...,k-1,

with the two boundary conditions:

(1) the first endslope

$$ s_0 = l'(1) = -(365/31)\,l(1)D_m/[B-D_1+D_m], \quad (5.3a) $$

(2) the last endslope

$$ s_k = l'(90) = -l(90)\,({}_5M_{85})^{3/2}\,({}_5M_{80})^{-1/2}. \quad (5.3b) $$

The above boundary conditions define a complete cubic spline and are obtained on the basis of properties of life table functions. For further details see Hsieh (1977).

In Table 2 we compare four methods of computing approximate values of L_i, using the same set of data $\{x_i, l_i\}$ taken from Makeham curve (4.2). The exact values of L_i are obtained by integrating $l(x)$ in (4.2) from x_i to x_{i+1}.

The other three methods are:

Keyfitz and Frauenthal:

$$ L_i = \frac{n(l_i-l_{i+1})}{\ln l_i - \ln l_{i+1}}\,[1+n(M_{i+1}-M_{i-1})/24] \quad (5.4) $$

Polynomial (cubic):

$$ L_i = (13/24)n_i(l_{i+1}+l_i) - n_i(l_{i+2}+l_{i-1})/24 \quad (5.5) $$

Simple ratio:

$$ L_i = (l_i-l_{i+1})/M_i. \quad (5.6) $$

The age specific death rates

$$ M_i \equiv {}_{n_i}M_{x_i} $$

in (5.4) and (5.6) are computed from (2.2) using (4.1) and (4.2).

The results obtained from (5.1) generate a cumulative absolute error of <u>114</u>, as compared with <u>677</u>, <u>1134</u> and <u>16347</u> for formulas (5.4),(5.5) and (5.6) respectively.

To illustrate the present proposed method, formulas (3.6) and (5.1) are used to construct an abridged life table for the 1970-72 Canadian male population as shown in Table 3.

ACKNOWLEDGEMENT

Grateful acknowledgement is made to Douglas M. Okamoto for his assistance in performing the tests and in computing the three tables.

REFERENCES

1. Greville, T.N.E. (1943),"Short Methods of Constructing Abridged Life Tables." Rec. Am. Inst. of Actuaries 32: 29-42.
2. Greville, T.N.E. (1947), United States Life Tables and Actuarial Tables, 1939-1941. Washington, D.C.: National Office of Vital Statistics 117-122.
3. Hsieh, J.J. (1976), "Methods of Computing Person Years." Presented at Annual Joint Statistical Meetings in Boston.
4. Hsieh, J.J. (1977). "A Method of Life Table Construction and Spline Interpolation." Submitted to the Journal of the American Statistical Association.
5. Keyfitz, N. and Frauenthal, J. (1975), "An Improved Life Table Method." Biometrics 31: 889-8
6. Reed, L.J. and Merrell, M. (1939), "A Short Method for Constructing An Abridged Life Table." Am. J. Hygiene 30: 33-62.