The Probabilistic Life Table and Its Applications

Nan Li, United Nations[1]
Shripad Tuljapurkar, Stanford University

Life-table variables are assigned certain values at all ages, and hence are treated as deterministic. Moreover, the starting number of the hypothetical cohort is taken arbitrarily as 100 thousand. But the survival process of the hypothetic cohort is uncertain unless the starting number is infinite, and the uncertainty depends on the starting number. Thus, the nature of life tables is probabilistic; and the starting number of the hypothetic cohort should not be arbitrary. This paper provides a method to compute the probabilistic life table, of which the starting number for the hypothetical cohort is the number of birth in a stationary population that is the closest to the observed population in the referred period. In a probabilistic life table, each variable has a probability distribution rather than a sample value, at all the ages except the first one. Using probabilistic life tables, one can tell, for example, whether the difference between two life expectancies is statistically significant or not. Finally, deterministic life tables are approximations for large populations.

In comparing mortality difference between countries or over times, the main difficulty comes from the effect of the population age structure. In a certain period, fewer deaths would occur in a population with younger age structure. But the age structure is determined by historical changes in fertility, mortality, and migration, and has nothing to do with the current mortality. To overcome this difficulty, a hypothetical cohort, which obeys the observed probabilities of death at corresponding ages, and is closed to migration, is proposed to reflect the effect of only mortality. To describe the survival process of this hypothetical cohort, variables at certain ages, such as the number of survivors, or the life expectancy, are constructed; and these variables compose a life table.

[1] Views expressed in this paper are solely those of the authors and do not necessarily reflect those of the United Nations.

Despite various ways of computing their variables, deterministic life tables have two common features. The first is that the starting number of the hypothetical cohort is arbitrary, such as 100 thousand. And the second is that all life-table variables are assigned certain values at all ages. It is obvious, however, that the survival process of a cohort is uncertain, unless the starting number of the cohort is infinite. It is also intuitive that the uncertainties depend on the starting number of the cohort. Thus, the two features should be removed.

In our previous work (Li and Tuljapurkar, 2012), the feature of arbitrarily taking the starting number of the hypothetical cohort has been removed, and life expectancy at birth has been extended from a deterministic variable to a probabilistic variable. This paper will provide a method to extend all life-table variables from deterministic to probabilistic, which leads to produce probabilistic life tables. This paper will also show the benefits of using probabilistic life tables.

The starting number of a probabilistic life table

As is mentioned above, a life table monitors the survival process of a hypothetical cohort that is closed to migration; and the starting number of the hypothetic cohort, l_0, is arbitrary in deterministic life tables. In the first life table compiled by John Grant in 1662 (see Pollard, 1973), l_0 was taken as 100 in the manner of giving examples. The value of l_0 is now commonly used as 100 thousand, still without a reason.

Does the value of l_0 matter? The old answer is it does not matter, in computing all life-table variables. For example, life expectancy at birth is computed as

$$e_0 = \frac{1}{l_0}\int_0^\infty l_a da = \frac{1}{l_0}\int_0^\infty l_0(1-{}_a q_0)da = \int_0^\infty (1-{}_a q_0)da, \qquad (1)$$

where l_a and ${}_aq_0$ represent the number of survivors at age *a* and the probability of dying between age 0 and *a*, respectively. The old answer is based on the assumption that all individuals in the hypothetic cohort survive according to the specified probabilities of death by age, which is taken as the first assumption of the paper. The second assumption of this paper, which was not required by the old answer, is that these individuals survive independently each other. Under the two assumptions, the survival process can be described by a binomial distribution:

$$l_a \sim B(l_0,\ 1-{}_aq_0),\ \ mean[\frac{l_a}{l_0}]=1-{}_aq_0,\ \ \mathrm{var}[\frac{l_a}{l_0}]=\frac{{}_aq_0(1-{}_aq_0)}{l_0}\ . \quad (2)$$

It can be seen that when the l_0 were infinitively large, there would be $\mathrm{var}[\frac{l_a}{l_0}]\approx 0$, and therefore the old answer would be correct. However, if a country's population is stationary and the life table refers to a calendar year, then the l_0 is the number of annual births, not infinitively large. Thus, a new answer is reached for the real situation that the size of population is not infinitively large: the value of l_0 does not matter only in computing the mean values, but matters in describing other aspects, of all life-table variables. According to this answer, the nature of life tables is probabilistic, and the deterministic life table should be extended to probabilistic.

For observed populations that are not stationary, Li and Tuljapurkar (2012) suggested estimate the stationary-equivalent population, whose age-specific numbers are the closest to that of the observed population. Of this stationary-equivalent population, the l_0 is

$$l_0=\frac{\sum_{a=0}^{\omega}{}_nL_a\sum_{a=0}^{\omega}p_o(a,a+n)_nL_a}{e_0\sum_{a=0}^{\omega}{}_nL_a^2}, \qquad (3)$$

where $p_o(a, a+n)$ and ${}_nL_a$ represent the person-years in age group [a,a+n) of the observed and the stationary population (with arbitrary l_0), respectively; and ω the lower bound of the open age group. Computing this l_0 does not require data additional to that of a life table. Hereafter in this paper, the value of l_0 is computed by (3), no longer 100 thousand or any other arbitrary number.

To reflect of the effects of the relatively small, middle[2], and large population sizes, data on female death rate and population by age of Iceland, Switzerland and Japan in 2005 are chosen from the Human Mortality Database (HMD, http://www.mortality.org/); and these are all the data used in this paper. The observed and stationary-equivalent populations are shown in Figure 1.

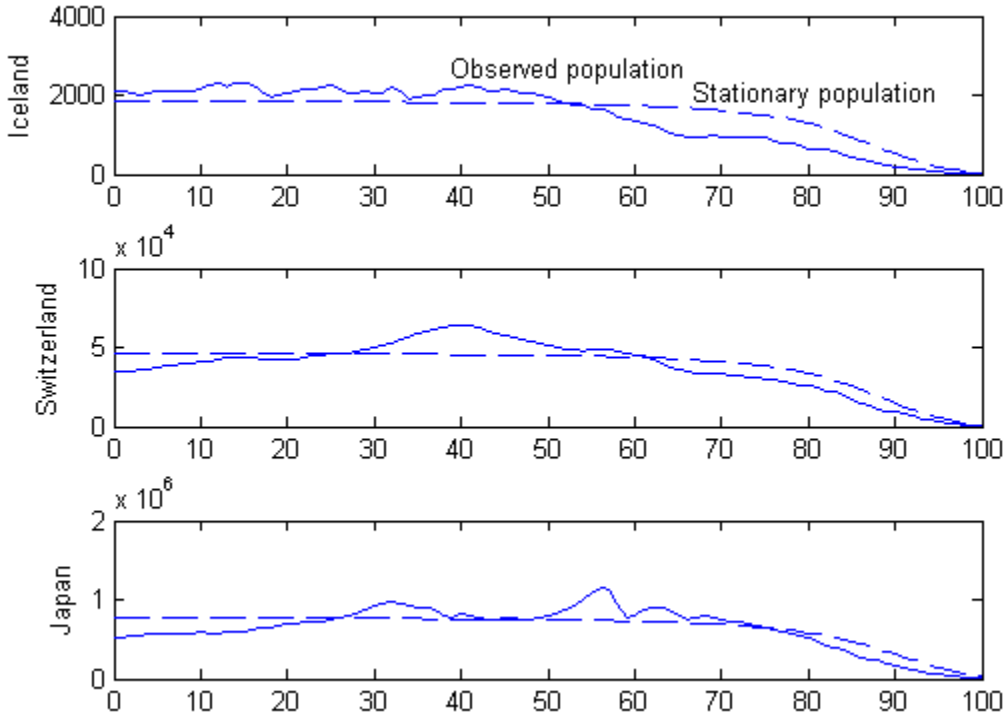

Figure 1. Observed and stationary female populations, 2005

[2] Among the 230 countries and areas of the world in 2005, the number of countries with female population larger than that of Switzerland was 93.

It can be seen that the values of l_0 are close to the numbers of annual births, but they differ remarkably between the small, median and large countries. The stationary populations are also the best cohort models, in the sense that they are the cohorts closest to the observed populations.

The definition of death probability and the measurement uncertainty of mortality

In statistics, the probability of dying between ages *a* and *a*+*n*, ${}_nq_a$, can be defined as

$$ {}_nq_a = \frac{l_a - l_{a+n}}{l_a}, \qquad (4) $$

when l_a is infinitively large (see Agresti and Finlay, 1997). In demography, the definition of death probability is often only (4), without requiring the condition of 'when l_a is infinitively large' (see Chiang, 1984). Comparing the two definitions, it is clear that the demographic definition is consistent with the old answer on how to choose l_0, while the statistics definition is the basis of the new answer. Ignoring the condition of 'when l_a is infinitively large ', we will miss the probabilistic nature of life tables. To discuss the measurement uncertainty of mortality, this condition is necessary, as will see below.

In order to understand and to use the statistics definition, we first discuss the probability of throwing a perfect coin and watching the face. The true value of this probability is, intuitively, 0.5. But throwing the coin once, or using sample size one, the observed value of the probability, computed as the number of having the face divided by the number of throwing, can only be either 0 or 1. Using large sample size, the observed value will converge to the true value, according to the large number law (see Agresti and Finlay, 1997). This is the reason of requiring the condition of 'infinitively large'.

We then turn to a more real situation that the true value of the probability is unknown, corresponds to throwing an imperfect coin. In this situation, the true value can still be defined as the observed value when the sample size approaches infinity, according to the discussion of the perfect coin.

We now reach the reality in which the true value is unknown and the sample size is not infinitively large. In reality, the true value can be described as the observed value plus a random variable, of which the mean is zero and the variance is smaller when the sample size is larger, according to (2).

Similarly, an observed probability of death, calculated from the data of death registration and population count, is a sample of death probability. Different from the case of throwing a coin, in which the chance for the face to appear is naturally random, the difference between the true and the observed values of death probability could be affected by random and avoidable errors. When the avoidable error does not exist, a sample is called random or unbiased. From this point of view, a sample of throwing a coin is naturally unbiased. A sample of death probability, however, may not be unbiased, because there could be avoidable errors of miscounting death and population. The third assumption of this paper is that the observed, or sample, probabilities of death by age are unbiased. Under this assumption, the true value of death probability can be defined as its sample value with infinitively large sample size. Moreover, when the sample size is not infinitively large, the true value of death probability can be expressed by a sample value plus a random variable, of which the mean is zero and the variance declines with the increase of sample size. The variance of this random variable could be computed according to (2), for some special cases such as ${}_aq_0$. This variance can also be computed numerically: generating a sample distribution of the death probability, and then calculate the variance using the sample distribution.

Similarly again, an observed life table, which is an output of the observed death probabilities at all ages, is a sample of the true life table. When the sample process, of death registration and population count, is unbiased, the true life table can be expressed

as the observed life table plus a random matrix, which describes the measurement uncertainty of mortality. Of this matrix, each element corresponds to a life-table variable at a certain age, the mean values of these elements are zero according to the unbiased assumption, and the variances of these elements are also expected to be smaller when the sample size is larger, as will be indicated by the examples below. The observed life table and the random matrix compose the probabilistic life table. The reason of not dealing with life-table variables individually, but a whole life table, is that life-table variables are correlated, as will discuss below.

The uncertainty of a life-table variable is the measurement uncertainty, which is the cause of sample errors that are the differences between the observed and the true values. The measurement uncertainty should be distinguished from the forecast uncertainty, which is derived from model errors that are the differences between the observed and the model values.

Computing probabilistic life tables

Assuming that the l_0 births survive independently according to the observed death probabilities at each age, sample distributions of all life-table variables at each age can be computed numerically. For life expectancy at birth, which is a specific life-table variable at a specific age, Li and Tuljapurkar (2012) indicated that its probability distribution is close to normal when the value of l_0 is large, and provided an analytic way to compute its variance. In general, however, the probability distributions of life-table variables at all ages are hard to derive analytically, especially when the number of corresponding event is not sufficiently large; and we discuss how to generate sample distributions for all life-table variables.

In a deterministic life table, all variables are functions of only age, which is denoted as *a*. In a probabilistic life table, all variables are random variables, and hence their values are functions of age and a sample number, of which "0" is reserved for the observation. More specifically, in this paper we use l_a to describe the number of survivors at

age *a*, which is a random variable; and $l_a(0)$ to represent the observed sample of l_a. Similarly, all life-table symbols are used to describe the corresponding random variables, and to present a sample value when an index is attached. The basic part of computing a probabilistic life table, in this paper, is to generate other samples of l_a, namely $l_a(i)$, where *i=1,2,3,..,* s_n; and s_n stands for the number of sample life tables. Using a large s_n, the sample distributions of l_a, and of all other life-table variables, can be computed.

The kernel of computing the sample distributions is the random survival. Let ${}_nq_a(0)$ be the observed probability of dying between age *a* and *a+n*, the survival process can be characterized by a random function $\Delta({}_nq_a(0))$ that takes value 1 to represent the survival of an individual from age *a* to age *a+n*, or 0 to describe the death of this individual between ages *a* and *a+n*. Under the two assumptions mentioned above, namely that each individual survives independently from the others according to the specified probability of death by age, the survival process obeys a Bernoulli distribution with parameter ${}_nq_a(0)$. The values of $\Delta({}_nq_a(0))$ can be generated by almost any computing software. For example, if ${}_1q_0(0) = 0.1$, then among 100 values of $\Delta({}_1q_0(0))$, 1 would appear approximately 90 times, indicating roughly 90 would survive to age 1 among 100 births. The exact number of $\Delta({}_1q_0(0)) = 1$, however, is uncertain. Further, the uncertainty accumulates over age. In other words, at older ages the number of survivors is more uncertain than at age 1. These uncertainties can be described by using multiple sample cohorts, each starts with the same number of births, and survives randomly according to the observed ${}_nq_a(0)$.

Denoting by $s(a, j)$ the survival status of the *jth* individual at age *a*, and using $s(a, j) = 1$ to represent alive at age *a* and $s(a, j) = 0$ deceased before age *a*, then the values of $s(a, j)$ can be computed as

$$s(a, j) = s(a-n, j)\Delta({}_nq_{a-n}(0)). \qquad (5)$$

For each individual, his or her $s(a, j)$ starts with 1 at age zero, and at some unpredictable age it drops to zero, meaning death.

At any age *a*, summing up all alive individuals produces the number of survivors of a cohort, and therefore the first sample value of l_a is obtained as:

$$l_a(1) = \sum_{j=1}^{l_0} s(a, j), \qquad (6)$$

where l_0 represents the number of stationary-equivalent births, as is discussed in the previous section.

Independently regenerating another set of $\Delta({}_nq_a(0))$ for each of the l_0 births at all the ages, and then using (5) and (6), the values of the second sample, $l_a(2)$, are computed. Repeating this procedure, the sample distribution of l_a is obtained as $l_a(i)$ for *i=1,2,…*, s_n, where s_n represents the number of sample cohorts. Since the $l_a(i)$ are computed using the unbiased ${}_nq_a(0)$ in a numerical way that is equivalent to (2), they are unbiased samples.

Given a unbiased sample $l_a(i)$, all other life-table variables of the *ith* sample cohort can be computed using the procedure of computing a deterministic life table. The procedure of computing deterministic life tables, as is indicated in the appendix, aims to minimize the errors of using discrete data to estimating the continuous survival process. Since these avoidable errors are minimized, the result life table could be assumed an unbiased sample of the true life table. Thus, the *ith* random sample of the probabilistic life table is generated. The difference, between this sample and the observed life table, is a sample of the random matrix mentioned above. Finally, a large number, namely s_n, of random sample life tables compose a probabilistic life table. In other words, a probabilistic life table is an extension of a deterministic life table, in which each variable at a certain age is extended from a sample value to a probability distribution. Moreover, this extension does not require data additional to that of a deterministic life table.

Illustrations

The number of sample cohorts, s_n, is taken as 1000 in this paper, fewer than which the sample distribution may be unsmooth, and more than which longer computing times are required. Figure 2 displays the 1000 curves of $l_a(i)$ for i=1,2, …, 1000, for the three chosen countries. First, all the curves start from the values of l_0 in Figure 1. Second, it can be seen that the survival process of a cohort is probabilistic and cannot be properly described in a deterministic way, as are shown by the case of Iceland, and perhaps also of Switzerland. This process, however, could be approximately described by a deterministic curve when the size of population is large enough, as is indicated by the case of Japan. Thus, as monitors of survival processes, life tables are naturally probabilistic, and deterministic life tables are approximations for large populations.

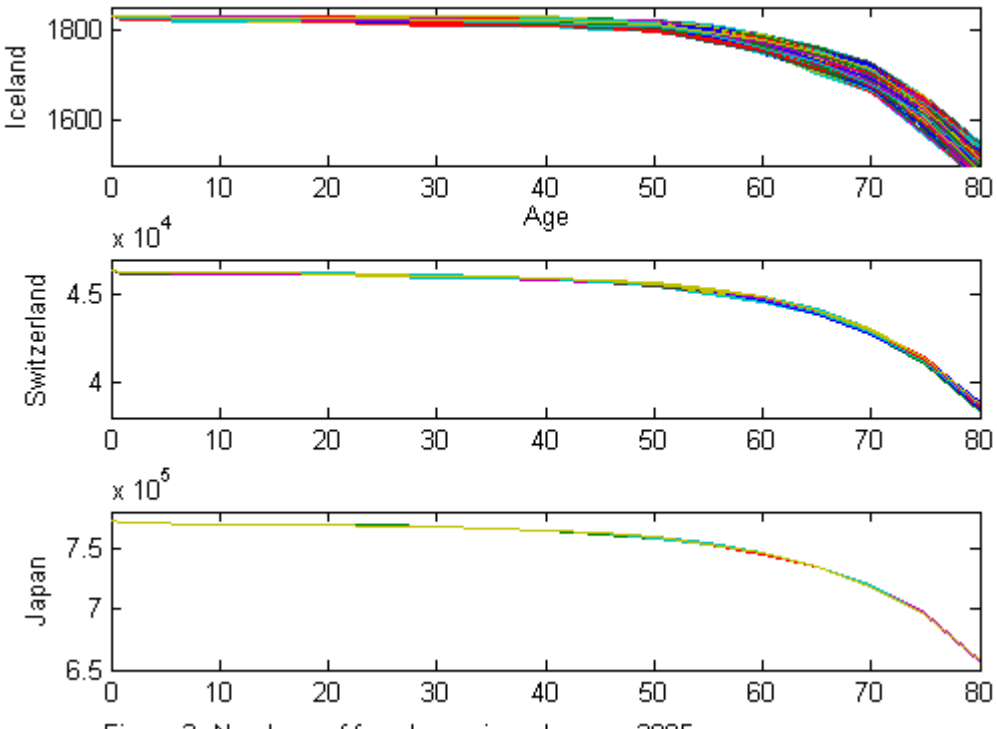

Figure 2. Numbers of female survivors by age, 2005

Furthermore, the *ith* sample, $l_a(i)$, produces *ith* sample of life table, which includes the *ith* sample of life expectancy at birth, $e_0(i)$. The 1000 samples of $e_0(i)$, subsequently, provide a sample distribution of life expectancy at birth, as is shown by the bars in Figure 3.

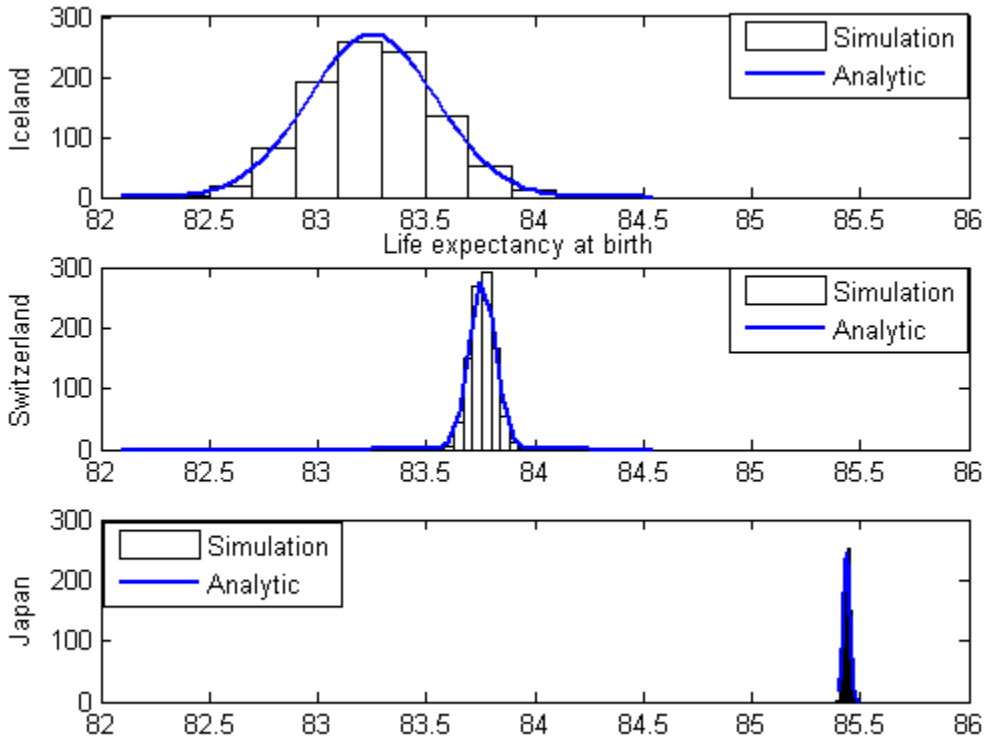

Figure 3. Sample distributions of life expectancy at birth, female 2005

The probability distributions of life expectancy were commonly computed by a method proposed by Chiang (1984), which is not based on the stationary population and therefore is logically improper. On the basis of the stationary population, Li and Tuljapurkar (2012) provided probability distributions of life expectancy at birth, as the curves show in Figure 3.

Applications based on one life table

Infant mortality rate (*IMR*), defined as the ratio of the number of deaths at ages younger than 1 year to the number of births in a certain period, is perhaps the most widely estimated mortality measure. The variance of *IMR* can be computed according to (2), using the observed number of annual births, namely b_o, as $IMR \cdot (1 - IMR) / b_o$, assuming that the deaths obey a binomial distribution. The observed *IMR* corresponds to the ${}_1q_0(0)$ in a life table, namely the probability of dying between birth and age 1, and the variance of ${}_1q_0$ is correspondingly ${}_1q_0(0) \cdot (1 - {}_1q_0(0)) / l_0$. Although it is often assumed that

${}_1q_0(0) = IMR$, the variances of ${}_1q_0$ and *IMR* can be different, because b_o is not exact l_0, which is computed by (3). In general, the observed period death rate is denoted as ${}_nM_a$ in order to be distinguished from the underlying cohort death rate ${}_nm_a(0)$. In the view of point estimates, the difference between ${}_nm_a(0)$ and ${}_nM_a$ is negligible because mortality could change only slightly in a moderate period, and it is therefore often assumed that ${}_nm_a(0)={}_nM_a$ (Preston, Heuveline, and Guillot, 2001). Similar to the case of ${}_1q_0(0)$ and *IMR*, the variances of ${}_nm_a$ and ${}_nM_a$ could differ notably. On the one hand, the variance of ${}_nM_a$ can be computed according to binomial distribution (Chiang, 1984) as ${}_nM_a(1-{}_nq_a(0))/p_o(a,a+n)$, where $p_o(a,a+n)$ is the number of observed population in age group *[a,a+n)*. On the other hand, the denominator of ${}_nm_a$ is a random variable. Thus, the variance of ${}_nm_a$ should not be calculated as ${}_nm_a(0)(1-{}_nq_a(0))/{}_nL_a(0)$ according to the formula of ${}_nM_a$, of which the denominator is a deterministic number. Instead, the variance of ${}_nm_a$ should be computed using its sample distribution, ${}_nm_a(i)$, as $\sum_{i=1}^{s_n}[{}_nm_a(i)-\sum_{k=1}^{s_n}{}_nm_a(k)/s_n]^2/(s_n-1)$.

The uncertainties of *IMR* could be remarkable as are shown in Figure 4, but they are not properly noticed. Ignoring these uncertainties, over-time fluctuations in infant mortality for Iceland, as a country with reliable data, can hardly be understood. Using these uncertainties, such fluctuations can be explained as differences between random samples.

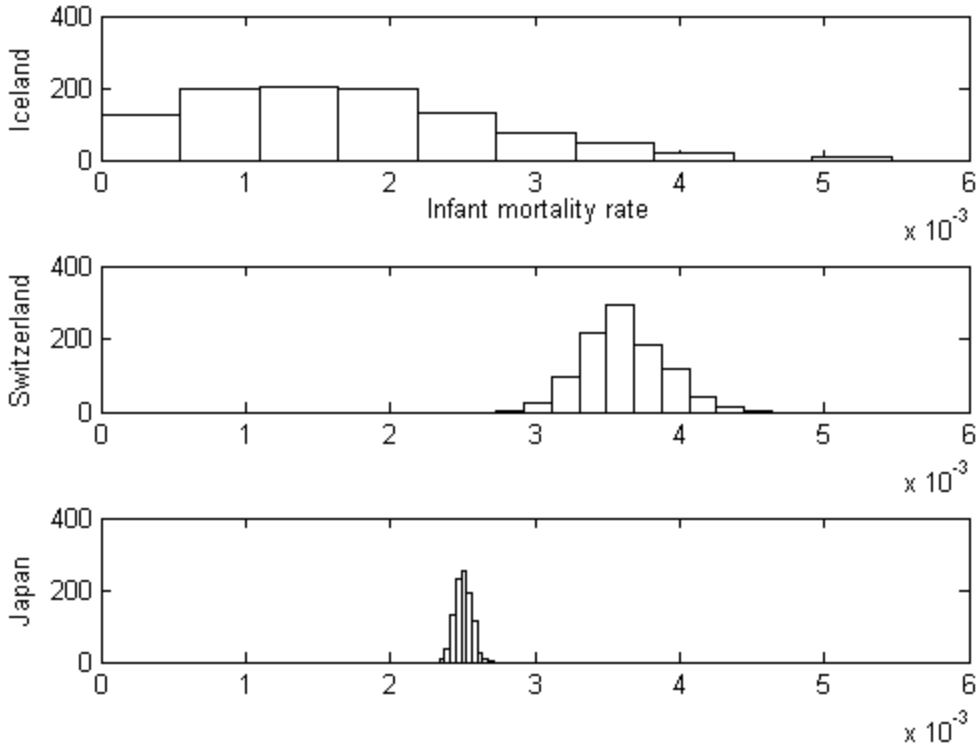

Figure 4. Sample distributions of infant mortality rate, female 2005

Adult mortality, usually defined as the probability of dying between ages 15 and 60 and denoted as ${}_{45}q_{15}$, is holding attention recently (e.g., Rajaratnam et al, 2010). There is no period correspondence of ${}_{45}q_{15}(0)$, and thus the variance of ${}_{45}q_{15}$ has to be computed in the context life table. Moreover, since the numerator and the denominator of ${}_{45}q_{15}$ are both random variables, the variance of ${}_{45}q_{15}$ should be computed using its sample distribution, ${}_{45}q_{15}(i)$, in the way same as that of ${}_{n}m_{a}(i)$ discussed above. Sample distributions of ${}_{45}q_{15}$ for the three countries are show in Figure 5.

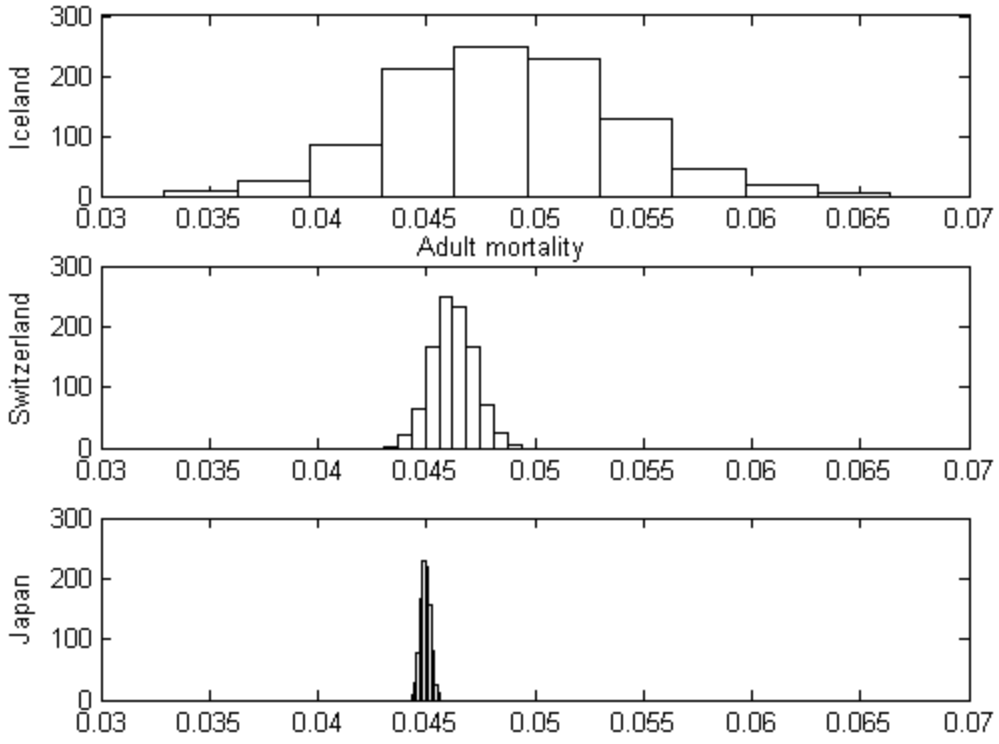

Figure 5. Sample distributins of adult mortality, female 2005

Comparing to infant and adult mortality, old-age mortality is perhaps more important, because it covers a larger fraction of deaths among a population. In this paper, we measure old-age mortality by the life expectancy at age 65, namely e_{65}. Because the numerator and the denominator of e_{65} are both random variables, the variance of e_{65} should not be computed by the formulas of e_0, but using the sample distribution, $e_{65}(i)$, which is displayed in Figure 6.

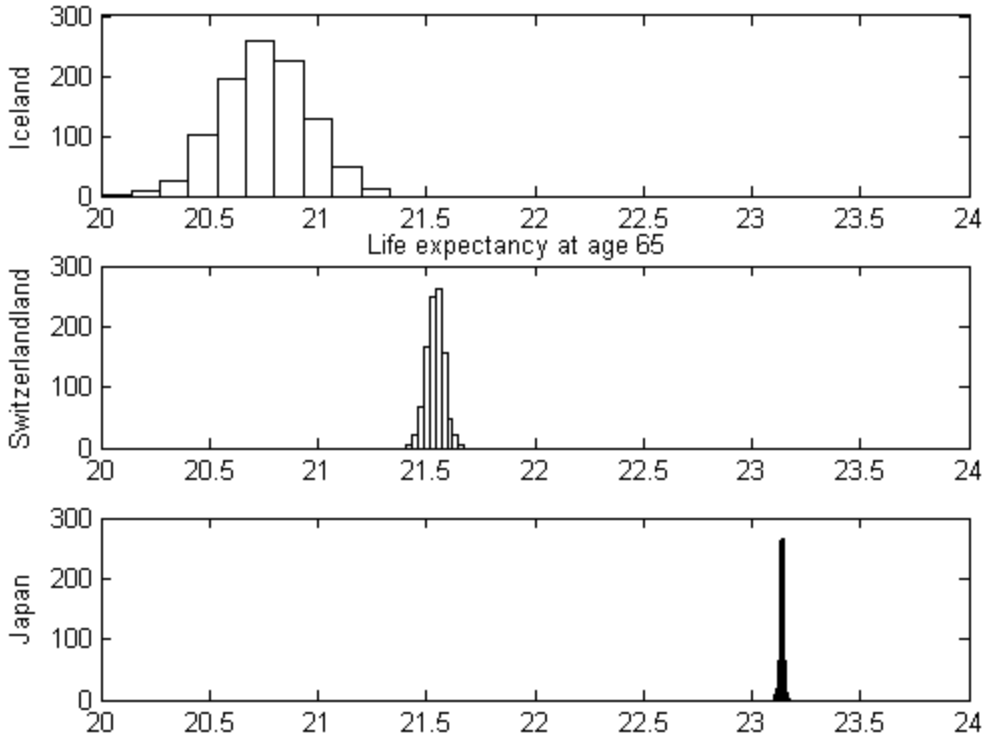

Figure 6. Sample distributions of life expectancy at age 65, female 2005

A practically important question is how to express a life-table variable. Taking life expectancy at birth as example, most countries publish its value using two decimals, but the United Nations Population Division (e.g., United Nations, 2009) uses one decimal, and the World Health Organization (e.g., WHO, 2005) shows only integer. In Figure 3, it can be seen that for Iceland, even the integer changes over the 95% confidence interval, and therefore using two decimals may be misleading. On the other hand, showing only integer for Japan is perhaps too rough. Thus, a certain number of decimals cannot fit all the countries; and the answer is beyond the format of a number. The applications based on a single life table indicate that, for most countries, a life-table variable can only be described properly by a probability distribution, not by a number of whatever formats.

Applications based on multiple life tables

The essential purpose of measuring mortality is to detect the difference between countries or the changes over times. To serve this purpose, deterministic life tables could

specify whether a life-table variable for one country or time is bigger than this variable for another country or time; while probabilistic life tables can further test whether such a difference is statistically significant, or it could appear merely by random chance.

Denote a life-table variable for populations 1 and 2 by x_1 and x_2, and the corresponding sample values by $x_1(i)$ and $x_2(i)$, *i=0,1,2,…* s_n. To test the significance of the difference between the mean values of x_1 and x_2, the null hypothesis can be set as *Ho:* $mean(x_1) = mean(x_2)$. Under this hypothesis, a test statistic can be constructed as

$$z = \frac{x_1 - x_2 - mean(x_1 - x_2)}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}, \qquad (7)$$

where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the estimated variances of x_1 and x_2, and can be computed from the sample distributions as

$$\sigma_h^2 = \sum_{i=1}^{s_n} [x_h(i) - \sum_{k=1}^{s_n} (x_h(k)/s_n]^2 / s_n, \quad h = 1,2. \qquad (8)$$

On the other hand, without the null hypothesis, the observed value of the test statistic *z* is

$$z(0) = \frac{x_1(0) - x_2(0)}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}, \qquad (9)$$

where the observed values of the life-table variable in question, namely $x_1(0)$ and $x_2(0)$, are computed from the corresponding deterministic life tables, in which, for example, $l_a(0) = l_0(1 - {}_a q_0(0))$.

In the usual applications of significance test, the analytic distribution of *z* is known (such as *N(1,0)*) and so does its 95% confidence interval, namely $[c_1, c_2]$ (such as

[-1.96,1.96]). If *z(0)* falls outside $[c_1, c_2]$, then the *Ho* is rejected at 0.05 level; and the conclusion is that the difference between the mean values of x_1 and x_2 is statistically significant. Otherwise, the difference between the means of x_1 and x_2 cannot be claimed statistically significant.

Here we do not know the analytic distribution, but we have the sample distribution, of *z*. Since the rank *i* for $x_1(i)$ and $x_2(i)$ are chosen randomly, the sample distribution of *z* can be constructed as

$$z(i) = \frac{x_1(i) - x_2(i) - \sum_{i=1}^{s_n} (x_1(i) - x_2(i))/s_n}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}, \quad i = 1 \sim s_n. \qquad (10)$$

Using the sample distribution of *z*, we can find out the 95% confidence interval, $[c_1, c_2]$, not exactly, but approximately. Subsequently, significance tests can be carried out, following the same logic of using analytic distributions.

Using this test to the life expectancy at birth of the three countries, results are shown in Figure 7 below

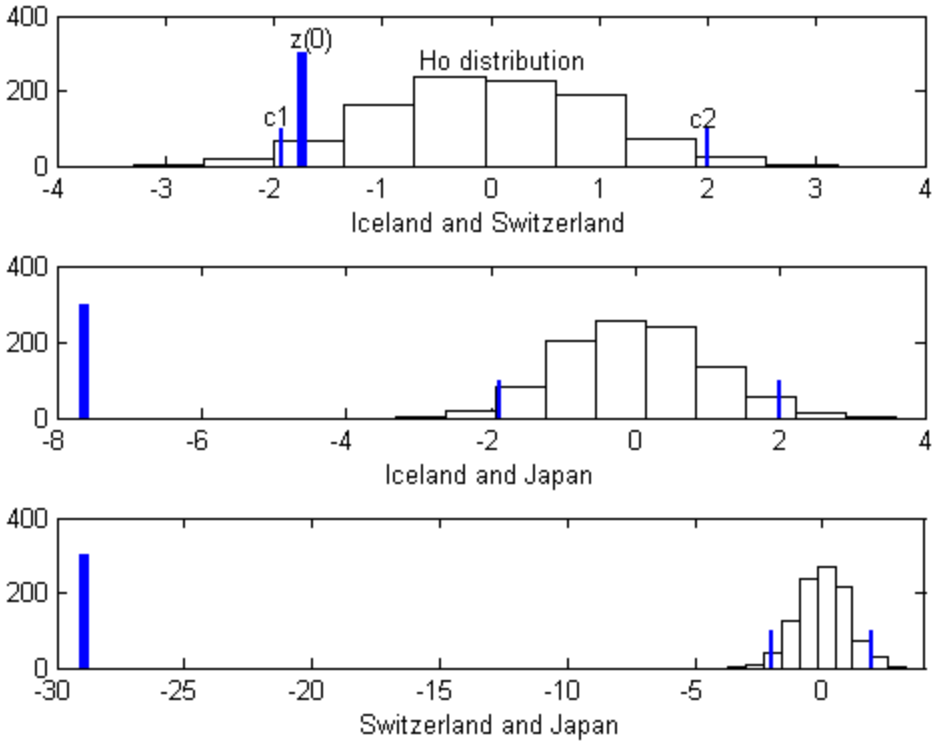

Figure 7. Significance test: difference between life expectancy at birth, female 2005

The probability distribution of the difference between life expectancies at birth is know as approximately normal (Li and Tuljapurkar, 2012), and the 95% confidence interval is $[c_1, c_2] \approx [-1.96, 1.96]$. It can be seen that the 95% confidence intervals obtained from the sample distributions are close to [-1.96, 1.96]. It can also be seen that the difference between the mean values of life expectancy at birth of Japan and the other two countries is significant; but between Iceland and Switzerland is not. This conclusion is consistent with the impression of Figure 3. Life expectancies at birth are compared for the purposes of reflecting the difference in development levels, but whether or not they are statistically significant had been rarely asked.

Applying the test to infant mortality rate, results are displayed in Figure 8. It can be seen that only the difference between Switzerland and Japan is significant, while the others are not. Infant mortality rates are compared to reflect the difference between health

care systems, and again whether or not they are statistically significant had hardly been noticed.

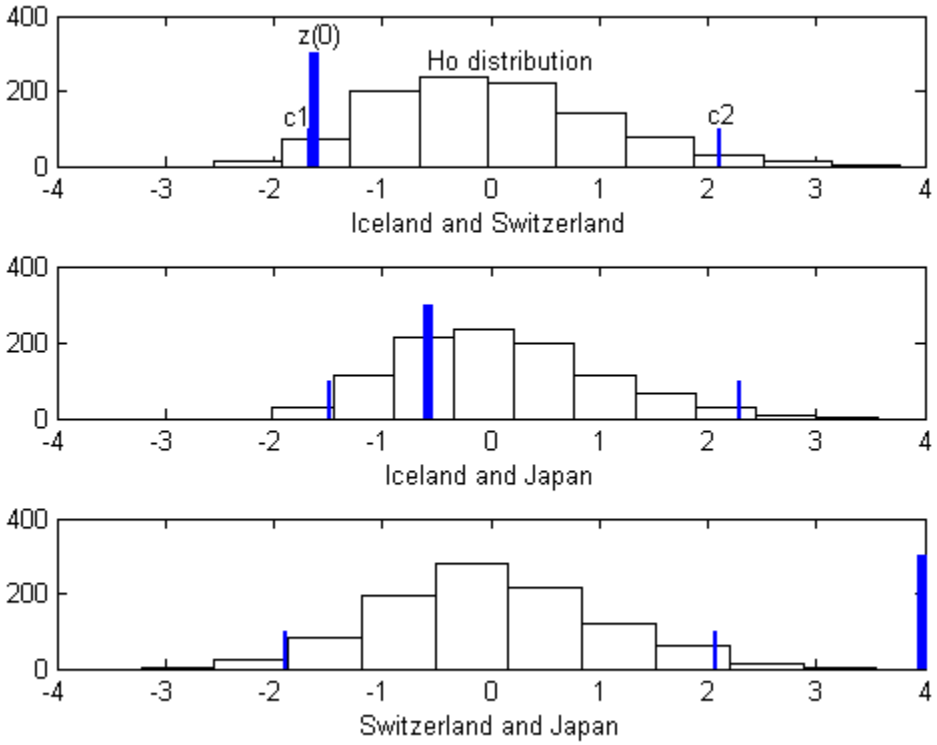

Figure 8. Significance test: difference between infant mortality rate, female 2005

Using the test on adult mortality, results are shown in Figure 9. It can be seen that, between the three countries, nor difference is statistically significant. This conclusion is somewhat surprising, but it is consistent with the intuitive impression from Figure 5.

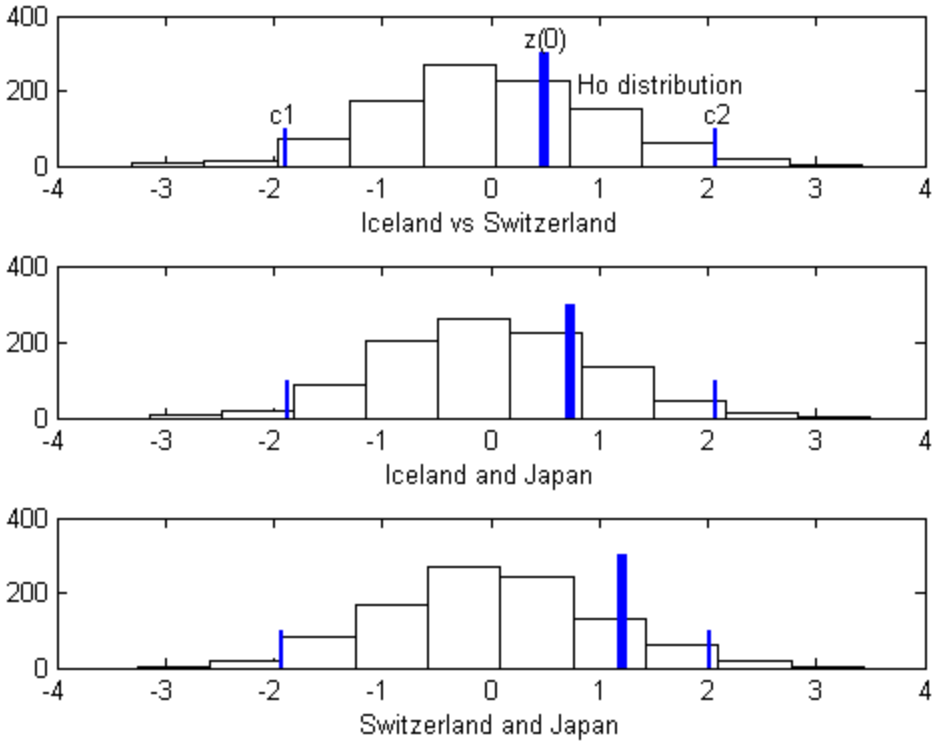

Figure 9. Significance test: difference between adult mortality, female 2005

Finally, applying the test to life expectancy at age 65, results are displayed in Figure 10. Entirely different from the difference between adult mortality, the differences between life expectancies at age 65 are all significant. This conclusion is also impressive, and it is consistent with Figure 6.

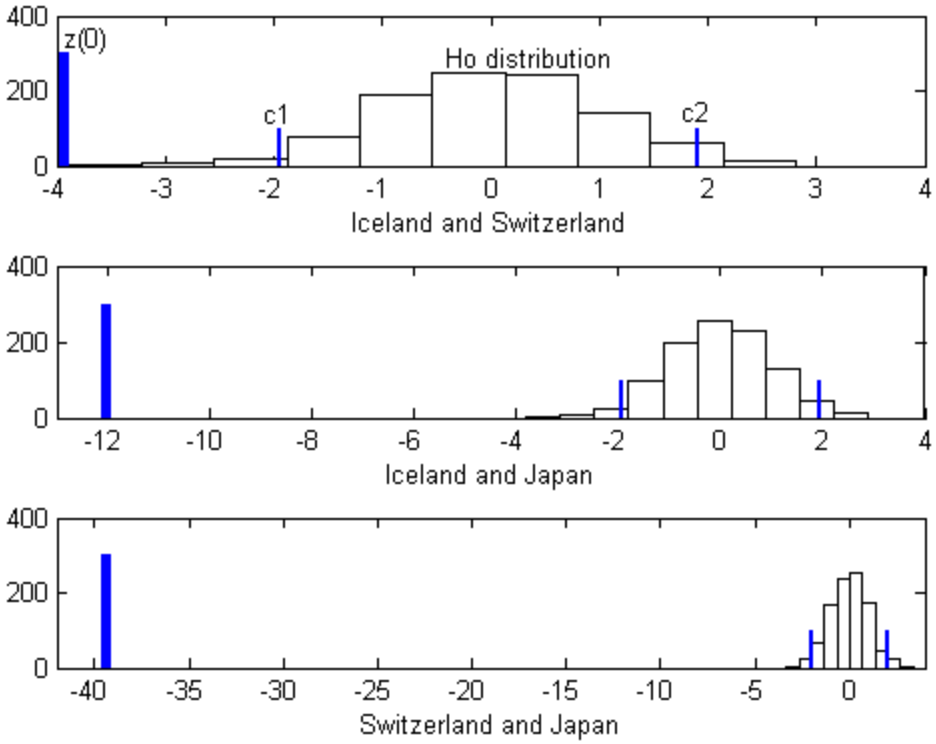

Figure 10. Significance test: difference between life expectancy at age 65, female 2005

Discussion

One may notice that, in Figures 3-10, the sample distributions are all similar to normal, and in Figures 7-10 the $[c_1, c_2]$ ranges are all close to [-1.96, 1.96]. These similarities cannot be coincidence, and they can be explained by the Central Limit Theorem (see Agresti and Finlay, 1997).

Let $Y(i, j)$ be a random variable assigned to the *jth* individual in the *ith* sample cohort, where $j=1,2,\ldots, l_0$ and i=1,2,…, s_n. Defining $Y(i, j)$ properly, the above four life-table variables for the *ith* cohort can be constructed as

$$X(i) = \frac{\sum_{j=1}^{l_a(i)} Y(i, j)}{l_a(i)}. \qquad (11)$$

For example, if $Y(i, j)$ is defined as the age of death of the *jth* individual who is alive at age *a=65*, then $X(i)$ is the life expectancy at age 65 of the *ith* sample cohort. For another example of *a=15* in $l_a(i)$, if $Y(i, j)$ is defined as

$$Y(i, j) = \begin{cases} 1, & \text{if} \quad S(15, j) = 1 \quad \text{and} \quad S(60, j) = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

where *S(a,j)=1 or 0* indicates that the *jth* individual is alive or deceased at age *a*, then the $[1 - X(i)]$ is the adult mortality of the *ith* sample cohort. For the same reason, the infant mortality rate and life expectancy at birth can also be presented by (11). Further, assuming that $Y(i, j)$ and $Y(i, k)$ are independently and identically distributed variables, then the $X(i)$, regardless of presenting which of the four life-table variables, is approximately a normal variable when $l_a(i)$ is large, according to the Central Limit Theorem. Since all the four variables of each of the sample cohort obey approximately normal distribution, so should be sample distributions in Figures 3-6.

Applying (10) and (11) to a life-table variable for two populations, namely x_1 and x_2 that can be assumed independent each other, then the *z* in (7) is not only a random variable with mean 0 and variance 1, but also approximately normal, according to the Central Limit Theorem. Thus, why all the sample distributions look normal and why all the ranges $[c_1, c_2]$ are close to [-1.96, 1.96], in Figures 7-10, are answered.

The number of stationary-equivalent births, l_0, and the number of sample cohorts, s_n, are essential in this paper. When the value of l_0 is larger, the variances of life-table variables will be smaller. On the other hand, when the value of s_n is larger, the sample distributions of these variables will be smoother, and the estimated variances from the sample distribution will be more accurate.

Summary

The age of death is uncertain for an individual. Accordingly, the nature of life tables is probabilistic. In this paper, we extended the life table from deterministic to probabilistic using three assumptions. The first assumption is that all individuals survive according to the specified probability of death by age, which is also the basis of the deterministic life table. The second assumption is that all individuals survive independently each other; and the third is that the observed probabilities of death by age are unbiased. Our extension requires two additional (the second and third above) assumptions, but not additional data. In a probabilistic life table, each variable has a probability distribution rather than a sample value, at all the ages except the first one.

Using applications of a single life table, we indicated that the uncertainties of life-table variables are too large to ignore. For most countries, a life-table variable at a certain age cannot be expressed properly by a number, regardless of using how many decimals; it can only be properly described by a probability distribution.

Using applications of multiple life tables, we showed that significance test can be carried out on all life-table variables. The difference between a certain variable in two life tables is essential in mortality and related studies. But mistakes could be made by explaining the socioeconomic reasons of a relatively large difference that may merely be a random effect, or by ignoring a relatively small difference that could turn out important. To avoid such mistakes, significance test can be helpful.

In this paper, the observed death probabilities, ${}_nq_a(0)$, are assumed unbiased, implying that there is no avoidable error of miscounting death or population. In our opinion, avoidable errors are negligible for most developed nations, and for some developing countries as well, because they have reliable vital registration and population census. For these countries, computing probabilistic life tables is a progress.

For many developing countries, however, the values of ${}_nq_a(0)$ are often obtained from model life tables and surveys focusing on only a few ages. Whether these ${}_nq_a(0)$ could be assumed unbiased, at all the ages, may be debatable. If the ${}_nq_a(0)$ could not be assumed unbiased, then how to apply the method of this paper will need further studies. The reason is, for example, a biased measure can make a difference statistically significant but it in fact is not. Nonetheless, data collections are improving among developing countries, which will make computing probabilistic life tables useful in the future.

As is shown above, the probabilistic life table improves mortality measure, but can it be useful? It depends on what computing power and carrying media are available. Using abacus for calculation and printing results on papers, the probabilistic life table cannot even exist. But the probabilistic life table can be useful, when calculations are done by computers and results are published on internet.

Appendix

This appendix focuses on the procedure of using the values of l_a to compute all other life-table variables for life tables, and the formulas apply to any sample $l_a(i)$. We first discuss the case of abridged life tables, of which age *a* takes the values of 0, 1, 5, 10, and so on until an open age group.

Given l_a, the number of death ${}_nd_a$ and the values of ${}_nq_a$ are immediately available:

$$\begin{aligned} {}_nd_a &= l_a - l_{a+n}, \\ {}_nq_a &= \frac{{}_nd_a}{l_a}. \end{aligned} \quad . \qquad \text{(A.1)}$$

For age groups 0-1 and 1-4, the calculations are based on the results of the West Family of the Coale-Demeny (1966) model life tables. Death rate ${}_1m_0$ and the person years lived by the deaths before age 1, ${}_1a_0$, are obtained as

$$ {}_1m_0 = \begin{cases} \dfrac{{}_1q_0}{1-(1-b_0)\cdot {}_1q_0}, \ \ if \ \ \dfrac{{}_1q_0}{1-(1-b_0)\cdot {}_1q_0} = A \geq 0.107, \\ \dfrac{-[1-(1-c_{00})\cdot {}_1q_0] + \sqrt{[1-(1-c_{00})\cdot {}_1q_0]^2 + 4\cdot c_{01}\cdot {}_1q_0^2}}{2\cdot c_{01}\cdot {}_1q_0}, \ \ if \ \ A < 0.107, \end{cases} \quad \text{(A.2)} $$

$$ {}_1a_0 = \frac{{}_1q_0 \cdot (1+{}_1m_0) - {}_1m_0}{{}_1q_0 \cdot {}_1m_0}. \quad \text{(A.3)} $$

And the values of ${}_4a_1$ and ${}_4m_1$ are computed as

$$ {}_4a_1 = \begin{cases} b_1, \ {}_1m_0 \geq 0.107, \\ c_{10} + c_{11}\cdot {}_1m_0, \ {}_1m_0 < 0.107, \end{cases} \quad \text{(A.4)} $$

$$ {}_4m_1 = \frac{{}_4q_1}{4-(4-{}_4a_1)\cdot {}_4q_1}. \quad \text{(A.5)} $$

The values of parameters in (A.2)-(A.5) are in Table A.1

Table A.1 Parameters of (2) and (3)

	b_0	b_1	c_{00}	c_{01}	c_{10}	c_{11}
Male	.33	1.352	.045	2.684	1.651	-2.816
Female	.35	1.351	.053	2.8	1.522	-1.518

For other age groups, the values of ${}_na_a$ and ${}_nm_a$ come from solve the below equations:

$$\begin{cases} {}_n a_a = \dfrac{n}{2} - \dfrac{n^2}{12}[{}_n m_a - \dfrac{{}_n m_a - {}_n m_{a-n}}{n \cdot {}_n m_a}], \\ {}_n m_a = \dfrac{{}_n q_a}{n - (n - {}_n a_a) \cdot {}_n q_a}, \end{cases} \qquad \text{(A.6)}$$

of which the first is the Greville (1943) formula and the second is a definition.

Having the values of ${}_n m_a$, the number of person-years lived by the cohort between ages a and $a+n$ is obtained as

$${}_n L_a = \frac{{}_n d_a}{{}_n m_a}. \qquad \text{(A.7)}$$

Finally, the other two life-table variables, namely the total person years lived over age a and the life expectancy at age a, can then be computed as

$$\begin{aligned} T_a &= \sum_{y=a} {}_n L_y, \\ e_a &= \frac{T_a}{l_a}. \end{aligned} \qquad \text{(A.8)}$$

Finally, for the open age group starting at age ω, ${}_\infty L_\omega = 1/{}_\infty m_\omega(0)$, assuming that the population is stationary in the open age group.

The above formulas apply to abridged life tables. For complete life tables, of which the length of all age intervals is one year, (A.1)-(A.3) still apply for age group [0,1), and so does ${}_\infty L_\omega = 1/{}_\infty m_\omega(0)$ for the open age group. For other age groups, one could follow the Human Mortality Database (www.mortality.org) to use ${}_1 a_a = 0.5$. Having l_a and ${}_1 a_a$, ${}_1 L_a$ can be computed by

$${}_1 L_a = {}_1 l_{a+1} + {}_1 a_a \cdot ({}_1 l_a - {}_1 l_{a+1}), \qquad \text{(A.9)}$$

and all other life-table variables can be calculated accordingly.

References

Agresti, A. and B. Finlay, 1997. Statistical Methods for the Social Sciences. Prentic Hall, Inc., Upper Saddle River, NJ 07458.

Coale, A.J. and P. Demeny, P., 1966. Regional Model Life Tables and Stable Populations. Princeton, N.J., Princeton University Press.

Chiang, C. L., 1984. The Life Table and Its Applications. Robert E. Krieger Publishing Company, Malabar, FL, USA.

Greville, T. N. E., 1943. Short methods of constructing abridged life tables. The Record of the American Institute of Actuaries, (32)29-43.

Li, N. and S. Tuljapurkar, 2012. On the Accuracy of Life Expectancy. Paper presented at the 2012 annual meeting of Population Association of America, San Francisco.

Pollard, J. H, 1973. Mathematical models for the growth of human population. Cambridge University Press, 1973.

Preston, S. H., P. Heuveline, and M. Guillot, 2001. *Demography, Measuring and Modeling Population Process*. Blackwell Publishers Ltd, Oxford, UK.

Rajaratnam, J.K., J.R. Marcus, A. Levin-Rector, A.N. Chalupka, H. Wang, L. Dwyer, M. Costa, A.D. Lopez, C.J.L. Murray, 2010. Worldwide mortality in men and women aged 15–59 years from 1970 to 2010: a systematic analysis. The Lancet. 2010 Apr 30; 375:1704–1720.

United Nations, 2009. World Population Prospects: The 2008 Revision. United Nations, New York, 2009.

WHO, 2005. The World Health Report 2005. World Health Organization, Geneva, Switzerland.