A STOCHASTIC STUDY OF THE LIFE TABLE AND ITS APPLICATIONS: I. PROBABILITY DISTRIBUTIONS OF THE BIOMETRIC FUNCTIONS[1]

CHIN LONG CHIANG

Division of Biostatistics, School of Public Health, University of California, Berkeley, California, U. S. A.

1. INTRODUCTION

The life table is one of the oldest, most useful, and best-known topics in the field of statistics. It has many applications in various areas of research where birth, death, and illness may take place. The earliest life tables date as far back as the seventeenth century; Halley's famous table for the City of Breslau, published in the year 1693 [9], already contained most of the columns in use today. The subject matter, however, is by no means limited to human beings. Zoologists, biologists, physicists, manufacturers, and investigators in other fields have found the life table a valuable means of presenting their data. In spite of its popularity in many research areas, the life table as a subject has yet to be systematically explored from a statistical point of view.

There are two forms of the life table in general use: the cohort (or generation) life table and the current life table. In its strictest form a cohort life table records the actual mortality experience of a given group of individuals over a period of time extending from birth until the death of the last member of the group. A current life table, on the other hand, considers the mortality experience of an entire population at one point in time. The purpose of this investigation is to present a stochastic view of the subject, taking random variation into consideration and treating all the biometric functions as random variables. The results of our study will be given in a series of papers. In the first paper probability distributions of the main biometric functions are presented and formulas are derived for the corresponding mathematical expectations, variances, and covariances. Some of the findings are by no means original, but they are included for the sake of completeness.

---

[1]Presented at the joint meeting of the American Statistical Association and the Biometric Society, ENAR, in Atlantic City, September 13, 1957, under the title, "An application of stochastic processes to the life table and standard error of age-adjusted rates" [3].

Although each of the biometric functions has the same meaning and the same probability distribution in the cohort life table as in the current life table, it is important for the study of their random variation to keep in mind the order in which these functions are computed. In the cohort life table the number of survivors and the number of deaths are measured directly in an actual population; thus they are the basic random variables from which the proportion of deaths and other columns are obtained. In the current life table, on the other hand, the column of the proportion of deaths is first computed from the population death rate; other biometric functions are random variables only because they are functions of this proportion. We shall, therefore, in the second paper of this series present formulas for the sample variances and covariances of the biometric functions in terms of the number of survivors for the cohort life table and in terms of the actual age specific mid-year population and age specific death rate for the current life table.

The third paper will be devoted to the application of these formulas to practical problems in follow-up studies of patients affected with specific diseases in which there are some survivors on the closing date of the study; because of incompleteness of information, expectation of life and some other quantities in the life table cannot then be computed by the conventional method. Here we suggest a convenient means of computing the observed expectation of life and the corresponding standard error. The problem of competing risks is also treated. An actual follow-up study will be used by way of illustration.

The general form of the life table is reproduced below for the purpose of reference; the symbols used deviate slightly from the conventional ones in order to simplify formulas in the text. For a detailed description of life-table structure, the reader is referred to the work of Dublin, Lotka, and Spiegelman [5], Greville [8], and Reed and Merrell [11].

In the table, and throughout this paper, the term "age" refers to the *exact* age. The symbol x_i is the age at the beginning of the interval i; x_w will be used to denote the age at the beginning of the final interval in any given life table.

The age x_0 may be taken as 0, the time of birth, and l_0 the size of the original cohort. From l_0 on, all the biometric functions in the above table are treated as random variables that are estimators of the corresponding unknown quantities. The symbol q_i will be used to denote the unknown true probability of a person of age x_i dying between x_i and x_{i+1}, and e_i the true expectation of life at age x_i, for $i = 0, 1, \cdots, w$.

The term "*observed* expectation of life" is introduced for the symbol $\hat{e}_i$ to distinguish it from its unknown true value e_i. Because of their

LIFE TABLE

Age interval (years)	Number of survivors at age x_i	Proportion of deaths within age interval (x_i, x_{i+1})	Number of deaths within age interval (x_i, x_{i+1})	Number of years lived within age interval (x_i, x_{i+1})	Total no. of years remaining to survivors at age x_i	Observed expectation of life at age x_i
x_0 to x_1	l_0	$\hat{q}_0$	d_0	L_0	T_0	$\hat{e}_0$
...	...	...	...	...	...	...
x_i to x_{i+1}	l_i	$\hat{q}_i$	d_i	L_i	T_i	$\hat{e}_i$
...	...	...	...	...	...	...
x_w to x_{w+1}	l_w	$\hat{q}_w$	d_w	L_w	T_w	$\hat{e}_w$

limited use, we shall not discuss the distribution of the quantities in the columns L_i and T_i. If desired, their distributions and formulas for expectations and variances can be obtained, respectively, from those of l_i and $\hat{e}_i$.

In the text the following symbols will also be introduced:

q_{ij} = Pr [an individual alive at age x_i will die in interval (x_i, x_j)],
p_{ij} = Pr [an individual alive at age x_i will survive to age x_j].

When $x_j = x_{i+1}$, we will drop the second subscript and write q_i for $q_{i,i+1}$ and p_i for $p_{i,i+1}$. Obviously, q and p are complementary. The corresponding estimators are denoted by

$$\hat{q}_{ij} = 1 - l_j/l_i\,, \quad \hat{p}_{ij} = l_j/l_i\,, \quad \hat{q}_i = 1 - l_{i+1}/l_i\,, \quad \text{and} \quad \hat{p}_i = l_{i+1}/l_i\,.$$

Finally, we will write n_i to denote the length of the interval i (i.e., $x_{i+1} - x_i = n_i$). When n is equal to one for each age interval, we have the "complete" life table.

Throughout this investigation, we shall assume a homogeneous population in which each individual is subject to the same force of mortality and in which the probability of death for one individual is not influenced by the death of any other individual in the group.

2. PROBABILITY DISTRIBUTION OF l_x, THE NUMBER OF SURVIVORS AT AGE x

In the usual life table the various biometric functions are given only for integral ages or at other discrete intervals. In the derivation of the distribution of survivors, however, it is more convenient to treat age as a continuous variable and to derive formulas for l_x, the number of individuals surviving the age interval $(0, x)$, for any positive value x.

The distribution of l_x may be obtained by different approaches. Perhaps the simplest is to consider the l_x survivors as the number of successes in l_0 independent and identical trials with a probability p_{0x} of surviving the interval $(0, x)$. It follows then that l_x is a binomial random variable. However, this approach by itself does not give the formula for the probability p_{0x}. The explicit formula for p_{0x} can be derived by the "pure death process" (see, for example, [6] and [1]), which we shall sketch below.

Let μ_x be the force of mortality acting upon each individual in the original cohort l_0, such that

$$\mu_x \Delta x + o(\Delta x) = \Pr[\text{an individual alive at age } x \text{ will die between ages } x \text{ and } x + \Delta x], \quad \text{for} \quad x \geqq 0, \tag{1}$$

where Δx stands for an infinitesimal time interval and $o(\Delta x)$ a quantity of a smaller order of magnitude than Δx. We are interested in the probability function of l_x, given that there are l_0 individuals alive at age 0:

$$P_{l_0,k}(0, x) = \Pr[1_x = k \mid l_0 \quad \text{at age} \quad 0]. \tag{2}$$

The standard procedure for obtaining this probability function is to derive an explicit form of the probability generating function defined as

$$G_{l_x}(t, x) = E(t^{l_x} \mid l_0) = \sum_{k=0}^{l_0} t^k P_{l_0,k}(0, x). \tag{3}$$

The derivatives of this one function provide a convenient way of computing all of the probabilities in (2), and the moments of l_x as well. Using the established procedure [6], we found

$$G_{l_x}(t, x) = \left[1 - \exp\left\{-\int_0^x \mu_\tau \, d\tau\right\} + t \exp\left\{-\int_0^x \mu_\tau \, d\tau\right\}\right]^{l_0}. \tag{4}$$

Substituting

$$p_{0x} = \exp\left\{-\int_0^x \mu_\tau \, d\tau\right\}, \quad \text{for} \quad x \geqq 0, \tag{5}$$

for the exponential function in (4) gives the generating function of the probability stated in (2):

$$G_{l_x}(t, x) = [1 - p_{0x} + tp_{0x}]^{l_0}, \quad \text{for} \quad x \geqq 0. \tag{6}$$

Formula (6) will be recognized as the generating function of a binomial random variable in l_0 independent and identical trials with the binomial probability p_{0x} as given by formula (5). For $x = x_i$, we have the probability that an individual will survive the age interval $(0, x_i)$,

$$p_{0i} = \exp\left\{-\int_0^{x_i} \mu_\tau \, d\tau\right\}, \quad \text{for} \quad i = 0, 1, \cdots, w, \tag{5A}$$

and the generating function for the survivors l_i,

$$G_{l_i}(t, x_i) = [1 - p_{0i} + tp_{0i}]^{l_0}, \quad \text{for} \quad i = 0, 1, \cdots, w. \tag{6A}$$

We are now in a position to use the binomial theorem to obtain the required probability function for l_i,

$$\Pr[l_i = k \mid l_0 \text{ at age } 0] = \frac{l_0!}{k!\,(l_0 - k)!} p_{0i}^k q_{0i}^{l_0 - k},$$
$$\text{for} \quad k = 0, 1, \cdots, l_0; \qquad i = 0, 1, \cdots, w, \tag{7}$$

the mathematical expectation,

$$E(l_i \mid l_0) = l_0 p_{0i}, \quad \text{for} \quad i = 0, 1, \cdots, w, \tag{8}$$

and the variance

$$\sigma^2_{l_i \mid l_0} = \text{n} l_0 p_{0i} q_{0i}, \quad \text{for} \quad i = 0, 1, \cdots, w, \tag{9}$$

with $p_{0i} + q_{0i} = 1$.

In general, the probability of surviving an age interval (x_i, x_j) is given by

$$p_{ij} = \exp\left\{-\int_{x_i}^{x_j} \mu_\tau \, d\tau\right\}, \quad \text{for} \quad i \leqq j; \qquad i, j = 0, 1, \cdots, w, \tag{10}$$

with the obvious relationship,

$$p_{\alpha j} = p_{\alpha i} p_{ij}, \quad \text{for} \quad \alpha \leqq i \leqq j; \qquad \alpha, i, j = 0, 1, \cdots, w. \tag{11}$$

The generating function for the conditional distribution of l_j given l_i is

$$G_{l_j \mid l_i}(t_j) = E(t_j^{l_j} \mid l_i) = (1 - p_{ij} + t_j p_{ij})^{l_i},$$
$$\text{for} \quad i < j; \qquad i, j = 0, 1, \cdots, w. \tag{12}$$

When $j = i + 1$, (12) becomes

$$G_{l_{i+1} \mid l_i}(t_{i+1}) = E(t_{i+1}^{l_{i+1}} \mid l_i) = (1 - p_i + t_{i+1} p_i)^{l_i},$$
$$\text{for} \quad i = 0, 1, \cdots, w - 1. \tag{13}$$

Although formula (12) holds whatever may be $x_i < x_j$, it is important to point out that the conditional probabilities of l_j relative to $l_0, l_1, \cdots, l_i$ are the same as those relative to l_i in the sense that for each k

$$\Pr[l_j = k \mid l_0, \cdots, l_i] = \Pr[l_j = k \mid l_i],$$
$$\text{for} \quad i < j; \qquad i, j = 0, 1, \cdots, w.$$

In other words, the sequence $l_1, l_2, \cdots, l_w$ is a Markov process ([6], p. 338). Thus we have

$$E(l_j \mid l_0, \cdots, l_i) = E(l_j \mid l_i), \quad \text{for} \quad i < j; \quad i, j = 0, 1, \cdots, w, \tag{14}$$

and also

$$E(t_j^{l_j} \mid l_0, \cdots, i_i) = E(t_j^{l_j} \mid l_i), \quad \text{for} \quad i < j; \quad i, j = 0, 1, \cdots, w. \tag{15}$$

3. JOINT DISTRIBUTION OF $l_1, \cdots, l_w$, THE NUMBERS OF SURVIVORS

Following the idea of the preceding section, we introduce the generating function of the joint probability distribution of $l_1, \cdots, l_w$:

$$G_{l_1,\cdots,l_w}(t_1, \cdots, t_w) = E(t_1^{l_1} \cdots t_w^{l_w} \mid l_0), \tag{16}$$

which uniquely determines the joint probability

$$\Pr[l_1 = k_1, \cdots, l_w = k_w \mid l_0 \quad \text{at age} \quad 0].$$

Using a procedure described previously ([2], pp. 84–85) we obtain

Lemma 1. *The survivors $l_1, \cdots, l_w$ in the life table form a random vector with components having the binomial distribution; the generating function of the joint distribution and the covariance between any two of the random variables are given, respectively, by*

$$\begin{aligned} G_{l_1,\cdots,l_w}(t_1, \cdots, t_w) = [1 - \{&p_{01}(1 - t_1) + p_{02}t_1(1 - t_2) \\ &+ p_{03}t_1t_2(1 - t_3) + \cdots + p_{0w}t_1t_2 \cdots t_{w-1}(1 - t_w)\}]^{l_0} \end{aligned} \tag{17}$$

and

$$\sigma_{l_i, l_j} = l_0 p_{0j}(1 - p_{0i}), \quad \text{for} \quad i \leq j; \qquad i, j = 0, 1, \cdots, w. \tag{18}$$

Proof of formula (17) follows from the identity

$$E[t_1^{l_1} \cdots t_{i+1}^{l_{i+1}}] = E[t_1^{l_1} \cdots t_i^{l_i} E\{t_{i+1}^{l_{i+1}} \mid l_0, \cdots, l_i\}] \tag{19}$$

and from formula (15). Combining (15) and (19), we can write

$$E[t_1^{l_1} \cdots t_{i+1}^{l_{i+1}}] = E[t_1^{l_1} \cdots t_i^{l_i} E\{t_{i+1}^{l_{i+1}} \mid l_i\}], \qquad \text{for} \quad i = 0, 1, \cdots, w - 1, \tag{20}$$

where the conditional expectation of the quantity inside the braces is the generating function of the conditional distribution of l_{i+1} given l_i with the explicit function as presented in (13). Formula (17) is obviously true for $w = 1$, since in this case (17) becomes

$$G_{l_1}(t_1) = [1 - p_{01}(1 - t_1)]^{l_0}, \tag{21}$$

which is identical to (6) for $x = x_1$. Now suppose (17) is true for $w - 1$, and we may write

$$E[t_1^{l_1} \cdots t_{w-1}^{l_{w-1}}] = [1 - \{p_{01}(1 - t_1) + p_{02}t_1(1 - t_2) + \cdots + p_{0,w-1}t_1 \cdots t_{w-2}(1 - t_{w-1})\}]^{l_0}, \tag{22}$$

we want to prove that (17) is true also for w. Using identity (20) for $i = w - 1$, generating function (16) may be written as

$$G_{l_1,\cdots,l_w}(t_1, \cdots, t_w) = E[t_1^{l_1} \cdots t_{w-1}^{l_{w-1}} E(t_w^{l_w} \mid l_{w-1})]. \tag{23}$$

Writing (13) for $i = w - 1$ and substituting in (23) give

$$\begin{aligned} G_{l_1,\cdots,l_w}(t_1, \cdots, t_w) &= E[t_1^{l_1} \cdots t_{w-1}^{l_{w-1}}\{1 - p_{w-1} + t_w p_{w-1}\}^{l_{w-1}}] \\ &= E[t_1^{l_1} \cdots t_{w-2}^{l_{w-2}} s_{w-1}^{l_{w-1}}], \end{aligned} \tag{24}$$

where

$$s_{w-1} = t_{w-1}(1 - p_{w-1} + t_w p_{w-1}) = t_{w-1}[1 - p_{w-1}(1 - t_w)]. \tag{25}$$

Because of formula (22), (24) becomes

$$[1 - \{p_{01}(1 - t_1) + p_{02}t_1(1 - t_2) + \cdots + p_{0,w-2}t_1t_2 \cdots t_{w-3}(1 - t_{w-2}) + p_{0,w-1}t_1t_2 \cdots t_{w-2}(1 - s_{w-1})\}]^{l_0}. \tag{26}$$

Now substituting (25) in the last term inside the braces,

$$\begin{aligned} &p_{0,w-1}t_1t_2 \cdots t_{w-2}(1 - s_{w-1}) \\ &\quad = p_{0,w-1}t_1t_2 \cdots t_{w-2}[1 - t_{w-1}\{1 - p_{w-1}(1 - t_w)\}] \\ &\quad = p_{0,w-1}t_1t_2 \cdots t_{w-2}(1 - t_{w-1}) + p_{0w}t_1t_2 \cdots t_{w-1}(1 - t_w), \end{aligned} \tag{27}$$

where p_{0w} is written for $p_{0,w-1}p_{w-1}$ [equation (11)]. Formula (26) thus becomes identical with the generating function (17), and the proof is completed.

Formula (18) can be proven by direct computation from the relation

$$\sigma_{l_i,l_j} = \frac{\partial^2}{\partial t_i\, \partial t_j} G \,\Big|_{t=1} - \left(\frac{\partial}{\partial t_i} G \,\Big|_{t=1}\right)\left(\frac{\partial}{\partial t_j} G \,\Big|_{t=1}\right),$$

where the symbol G is written for the generating function (17) and the partial derivatives are taken at the point $(t_1, \cdots, t_w) = (1, \cdots, 1)$. When $i = j$, formula (18) reduces to the formula for the variance of l_i [equation (9)].

The joint probability of the random variables $l_1, \cdots, l_w$ can now be obtained from (17) by differentiating with respect to the arguments. It turns out to be

$$\Pr\{l_1 = k_1, \cdots, l_w = k_w \mid l_0\} = \prod_{i=1}^{w} \frac{k_{i-1}!}{k_i!\,(k_{i-1} - k_i)!} p_{i-1}^{k_i}(1 - p_{i-1})^{k_{i-1}-k_i},$$

for $k_i = 0, 1, \cdots, k_{i-1}$, with $k_0 = l_0$.

4. JOINT PROBABILITY DISTRIBUTION OF $d_0, \cdots, d_w$, THE NUMBERS OF DEATHS

In a life table covering the entire life span of each individual in a given population, the sum of the deaths at all ages is equal to the size of the original cohort. Symbolically,

$$d_0 + d_1 + \cdots + d_w = l_0. \tag{28}$$

Each individual in the original cohort has a probability of dying in the interval (x_i, x_{i+1}), which is easily shown to be

$$p_{0i}q_i, \quad \text{for} \quad i = 0, \cdots, w; \tag{29}$$

for, if an individual at age 0 is to die between ages x_i and x_{i+1}, he must first survive the age interval $(0, x_i)$. The multiplication theorem implies (29). Since he is to die once and only once somewhere in the life span covered by the life table, the sum of the probabilities in (29) is unity; or

$$p_{00}q_0 + \cdots + p_{0w}q_w = 1,$$

where $p_{00} = 1$ and $q_w = 1$. Thus we have the well-known

Lemma 2. *The numbers of deaths, $d_0, \cdots, d_w$, in a life table have a multinomial distribution with the joint probability distribution*

$$\Pr[d_0 = \delta_0, \cdots, d_w = \delta_w] = \frac{l_0!}{\delta_0! \cdots \delta_w!}(p_{00}q_0)^{\delta_0} \cdots (p_{0w}q_w)^{\delta_w}; \tag{30}$$

expectation, variance, and covariance are given, respectively, by

$$E(d_i \mid l_0) = l_0 p_{0i} q_i, \quad \textit{for} \quad i = 0, \cdots, w; \tag{31}$$

$$\sigma_{d_i}^2 = l_0 p_{0i} q_i (1 - p_{0i} q_i), \quad \textit{for} \quad i = 0, \cdots, w; \tag{32}$$

and

$$\sigma_{d_i, d_j} = -l_0 p_{0i} q_i p_{0j} q_j, \quad \textit{for} \quad i \neq j; \quad i, j = 0, \cdots, w. \tag{33}$$

Remark 1: In the above discussion, age 0 was chosen only for simplicity of presentation. For any given age, say x_α, the numbers of deaths occurring in subsequent intervals also have a multinomial distribution with the total number of deaths equal to the number of survivors at age x_α. The probability that an individual alive at age

x_α will die in the interval (x_i, x_{i+1}) subsequent to x_α is given by

$$p_{\alpha i}q_i , \quad \text{for} \quad i = \alpha, \cdots, w. \tag{34}$$

It can be readily shown that the sum of the probabilities in (34) is unity but we shall not give the details here.

5. VARIANCE AND COVARIANCE OF $\hat{q}_i$, THE PROPORTION OF DEATHS IN THE AGE INTERVAL (x_i, x_{i+1})

The proportion of deaths occurring in an age interval is the ratio of two random variables

$$\hat{q}_i = \frac{l_i - l_{i+1}}{l_i}, \quad \text{for} \quad i = 0, 1, \cdots, w - 1. \tag{35}$$

Our interest in this section is to derive formulas for the expectation, variance, and covariance of these proportions.

It is convenient at this point to reintroduce the proportion of survivors in the age interval (x_i, x_{i+1}),

$$\hat{p}_i = \frac{l_{i+1}}{l_i}, \quad \text{for} \quad i = 0, 1, \cdots, w - 1. \tag{36}$$

Since

$$\hat{p}_i + \hat{q}_i = 1, \tag{37}$$

the mathematical expectation of the proportion of deaths is complementary to the expectation of the proportion of survivors. These proportions have the same formulas for the variance and covariance: $\sigma^2_{\hat{q}_i} = \sigma^2_{\hat{p}_i}$, and $\sigma_{\hat{q}_i.\hat{q}_j} = \sigma_{\hat{p}_i.\hat{p}_j}$, for $i, j = 0, 1, \cdots, w - 1$.

The generating function (13) shows that the conditional distribution of l_{i+1} given l_i is binomial and has the conditional expectation

$$E(l_{i+1} \mid l_i) = l_i p_i , \quad \text{for} \quad i = 0, 1, \cdots, w - 1. \tag{38}$$

From (38) we derive the expectation of $\hat{p}_i$,

$$E(\hat{p}_i) = E\left(\frac{l_{i+1}}{l_i}\right) = E\left[\frac{1}{l_i} E(l_{i+1} \mid l_i)\right]$$
$$= E\left[\frac{1}{l_i} l_i p_i\right] = p_i , \quad \text{for} \quad i = 0, 1, \cdots, w, \tag{39}$$

and hence the expectation of $\hat{q}_i$,

$$E(\hat{q}_i) = 1 - p_i = q_i , \quad \text{for} \quad i = 0, 1, \cdots, w. \tag{40}$$

It is interesting to note from formula (10) that the ratio of the expectation of survivors at the end of an interval to the expectation of survivors at the beginning of the interval is also equal to the probability of surviving the interval. Consequently,

$$E\left[\frac{l_{i+1}}{l_i}\right] = p_i = \frac{E(l_{i+1})}{E(l_i)}, \quad \text{for} \quad i = 0, 1, \cdots, w-1, \tag{41}$$

the expectation of the ratio of the two random variables, l_{i+1} to l_i, is equal to the ratio of the expectations, a relationship not necessarily true in general.

The variance of $\hat{p}_i$ (or $\hat{q}_i$) may be written in the form

$$\begin{aligned}\sigma^2_{\hat{p}_i} &= E\left[\frac{l^2_{i+1}}{l^2_i}\right] - p^2_i \\ &= E\left[\frac{1}{l^2_i} E(l^2_{i+1} \mid l_i)\right] - p^2_i,\end{aligned}$$

where

$$E[l^2_{i+1} \mid l_i] = l_i p_i (1 - p_i) + l^2_i p^2_i$$

is again obtained from the generating function (13). By substitution and collection of terms we have the formula for the variance,

$$\sigma^2_{\hat{p}_i} = E\left(\frac{1}{l_i}\right) p_i (1 - p_i), \quad \text{for} \quad i = 0, 1, \cdots, w. \tag{42}$$

When l_0 is large, formula (42) may be approximated by

$$\sigma^2_{\hat{p}_i} = \frac{1}{E(l_i)} p_i (1 - p_i), \quad \text{for} \quad i = 0, 1, \cdots, w. \tag{43}$$

The expectation of the reciprocal of l_i can be written as

$$E\left(\frac{1}{l_i}\right) = \frac{1}{E(l_i)}\left\{1 + \frac{\sigma^2_{l_i}}{[E(l_i)]^2} + o\left(\frac{\sigma^2_{l_i}}{[E(l_i)]^2}\right)\right\}, \tag{44}$$

where the second term inside the square braces is the relative-variance of l_i and the third term is a quantity of a smaller order of magnitude than the relative-variance. Using the formulas (8) and (9) for the expectation and variance of l_i, we have

$$\frac{\sigma^2_{l_i}}{[E(l_i)]^2} = \frac{q_{0i}}{l_0 p_{0i}},$$

which may be taken as zero for large values of l_0. Consequently, the quantity inside the square braces in (44) may be taken as unity and formula (42) is approximated by (43)[2].

[2]It is obvious from formulas (35) and (36) that $\hat{q}_i$ and $\hat{p}_i$ are defined only for positive values of l_i. If l_i were equal to zero, d_i and l_{i+1} would certainly equal zero, and the biometric functions described in the life table, as well as the life table itself, will have ceased to be meaningful. Thus we shall use the convention that the denominator of (42) cannot take on the value of zero before the interval w_{+1}, which is to say, before the termination of the life table.

To derive the formula for the covariance between the proportions of survivors (or deaths) in two age intervals, we write

$$\sigma_{\hat{p}_i,\hat{p}_j} = E[\hat{p}_i\hat{p}_j] - p_ip_j$$
$$= E[\hat{p}_iE(\hat{p}_j \mid \hat{p}_i)] - p_ip_j\ , \quad \text{for} \quad i < j; \quad i, j = 0, 1, \cdots, w, \tag{45}$$

with the conditional expectation,

$$E(\hat{p}_j \mid \hat{p}_i) = E\left[\frac{l_{j+1}}{l_j} \middle| \frac{l_{i+1}}{l_i}\right].$$

Recalling from formula (14) that the conditional expectation of l_{j+1} relative to l_i , l_{i+1} , and l_j is the same as the conditional expectation of l_{j+1} relative to l_j , we have

$$E\left[\frac{l_{j+1}}{l_j} \middle| \frac{l_{i+1}}{l_i}\right] = E\left[\frac{1}{l_j} E(l_{j+1} \mid l_j) \middle| \frac{l_{i+1}}{l_i}\right]$$
$$= E\left[\frac{1}{l_j} l_jp_j \middle| \frac{l_{i+1}}{l_i}\right] = p_j\ . \tag{46}$$

Substitution of (46) in (45) gives the covariance

$$\sigma_{\hat{p}_i,\hat{p}_j} = E[\hat{p}_ip_j] - p_ip_j = p_ip_j - p_ip_j = 0,$$
$$\text{for} \quad i \neq j; \qquad i, j = 0, 1, \cdots, w. \tag{47}$$

Remark 2: What is proved above is the zero covariance between $\hat{p}_i$ and $\hat{p}_j$, but not their independence. In fact, it can be shown [4] that $\hat{p}_i$ and $\hat{p}_j$ are not independently distributed; and in particular, Greenwood's assumption [7] $E(\hat{p}_i^2\hat{p}_j^2) = E(\hat{p}_i^2)E(\hat{p}_j^2)$ is proven to be false.

The findings in this section may be summarized in

Lemma 3. *The proportion of deaths, $\hat{q}_i$, (or of survivors, $\hat{p}_i$) in an age interval is an unbiased estimator of the probability of dying in (or of surviving) the interval with a variance as given by* (42); *the covariance between two proportions $\hat{q}_i$ and $\hat{q}_j$ (or between $\hat{p}_i$ and $\hat{p}_j$) vanishes whatever may be $i \neq j$, for $i, j = 0, \cdots, w$.*

It should be pointed out that formula (47) of zero covariance is obtained only between proportions for two non-overlapping age intervals. If we are considering two intervals both beginning with the same age x_α and extending to the ages x_i and x_j , respectively, the covariance between the proportions $\hat{p}_{\alpha i}$ and $\hat{p}_{\alpha j}$ is not equal to zero. Using the same approach as in the derivation of (42), it is easy to show that the formula for the covariance is given by

$$\sigma_{\hat{p}_{\alpha i}, \hat{p}_{\alpha j}} = E\left(\frac{1}{l_\alpha}\right) p_{\alpha j}(1 - p_{\alpha i}), \quad \text{for } \alpha < i \leqq j; \quad \alpha, i, j = 0, \cdots, w. \tag{48}$$

When l_0 is large, we have the approximate formula

$$\sigma_{\hat{p}_{\alpha i}, \hat{p}_{\alpha j}} = \frac{1}{E(l_\alpha)} p_{\alpha j}(1 - p_{\alpha i}), \quad \text{for } \alpha < i \leqq j; \quad \alpha, i, j = 0, \cdots, w. \tag{49}$$

For $i = j$, (48) and (49) become formulas for the variance of $\hat{p}_{\alpha i}$. If $x_\alpha = 0$, l_0 is constant; both formulas (48) and (49) are reduced to

$$\sigma_{\hat{p}_{0i}, \hat{p}_{0j}} = \frac{1}{l_0} p_{0j}(1 - p_{0i}), \quad \text{for} \quad i \leqq j; \qquad i, j = 1, \cdots, w,$$

which is the covariance between the x_i- and x_j-year survival rates, and can be obtained directly from the covariance between l_i and l_j as given by (18).

6. DISTRIBUTION OF $\hat{e}_\alpha$, THE OBSERVED EXPECTATION OF LIFE AT AGE x_α

The observed expectation of life at any age x_i summarizes the mortality experience of the population under consideration beginning with age x_i, for $i = 0, 1, \cdots, w$. Certainly to the demographer or public health worker, this column is the most useful in the life table.

To avoid confusion in notation, let us denote by α a fixed number and by x_α a particular age; we are interested in the distribution of $\hat{e}_\alpha$, the observed expectation of life at the age x_α. Consider l_α, the survivors to the age x_α, and let Y_α denote the future lifetime of a particular individual beyond the age x_α. Clearly Y_α is a continuous random variable that can assume any positive real value. Let y_α be the value that the random variable Y_α takes on; thus $x_\alpha + y_\alpha$ is the entire length of life of the individual from the time of birth until death. Let $f(y_\alpha)$ be the probability density function of the random variable Y_α and dy_α an infinitesimal time interval. Since Y_α can assume values between y_α and $y_\alpha + dy_\alpha$ if and only if the individual of age x_α survives the age interval $(x_\alpha, x_\alpha + y_\alpha)$ and then dies in the interval $(x_\alpha + y_\alpha, x_\alpha + y_\alpha + dy_\alpha)$, the probability density function of Y_α is given by

$$f(y_\alpha)\, dy_\alpha = p_{\alpha, \alpha+y_\alpha} \mu_{x_\alpha + y_\alpha}\, dy_\alpha, \quad \text{for} \quad y_\alpha \geqq 0, \tag{50}$$

where $p_{\alpha, \alpha+y_\alpha}$, the probability of surviving the interval $(x_\alpha, x_\alpha + y_\alpha)$, is defined in (10) and $\mu_{x_\alpha + y_\alpha}$ is the force of mortality at age $x_\alpha + y_\alpha$ given in (1).

The function $f(y_\alpha)$ in (50) is an honest probability density function in the sense that it is never negative and that the integral of the function from $y_\alpha = 0$ to $y_\alpha = \infty$ is equal to unity. Clearly, it can never be negative, whatever may be the value of y_α. To evaluate the integral, we recall formula (10) and write

$$\int_0^\infty f(y_\alpha)\,dy_\alpha = \int_0^\infty \exp\left\{-\int_{x_\alpha}^{x_\alpha+y_\alpha} \mu_\tau\,d\tau\right\}\mu_{x_\alpha+y_\alpha}\,dy_\alpha\,.$$

Now define a quantity ϕ such that

$$\phi = \int_{x_\alpha}^{x_\alpha+y_\alpha} \mu_\tau\,d\tau = \int_0^{y_\alpha} \mu_{x_\alpha+t}\,dt$$

and substitute the differential

$$d\phi = \mu_{x_\alpha+y_\alpha}\,dy_\alpha$$

in the integral to give the solution

$$\int_0^\infty f(y_\alpha)\,dy_\alpha = \int_0^\infty e^{-\phi}\,d\phi = 1.$$

The mathematical expectation of the random variable Y_α is the expected length of future life beyond the age x_α, and thus may be called the true expectation of life at age x_α. In accordance with the definition given the symbol e_α, we may write

$$e_\alpha = \int_0^\infty y_\alpha f(y_\alpha)\,dy_\alpha = \int_0^\infty y_\alpha \exp\left\{-\int_{x_\alpha}^{x_\alpha+y_\alpha} \mu_\tau\,d\tau\right\}\mu_{x_\alpha+y_\alpha}\,dy_\alpha\,. \qquad (51)$$

Thus the explicit function of e_α and the variance of Y_α,

$$\sigma_{y_\alpha}^2 = \int_0^\infty (y_\alpha - e_\alpha)^2 \exp\left\{-\int_{x_\alpha}^{x_\alpha+y_\alpha} \mu_\tau\,d\tau\right\}\mu_{x_\alpha+y_\alpha}\,dy_\alpha\,, \qquad (52)$$

both depend on the force of mortality[3].

We will consider the future lifetimes of l_α survivors as a sample of l_α independent and identical random variables, $Y_{\alpha 1}, \cdots, Y_{\alpha l_\alpha}$, each of which has the probability density function (50), the mathematical expectation (51), and variance (52). According to the central limit theorem, as l_α approaches infinity, the distribution of the sample mean

$$\bar{Y}_\alpha = \frac{1}{l_\alpha}(Y_{\alpha 1} + \cdots + Y_{\alpha l_\alpha})$$

[3]While it is not the purpose of this paper to consider particular functions of the force of mortality, a separate study of the observed expectation of life under various assumption of the force of mortality is in preparation.

is approximately normal, with a mean of e_α as given in (51). Clearly $\bar{Y}_\alpha$ is equal to $\hat{e}_\alpha$, the observed expectation of life at age x_α.

As in the case of any continuous random variable, the value of Y_α is not accurately measured. In point of fact, the values of l_α random variables are not individually recorded in the life table, but rather they are grouped in the form of a frequency table in which the ages x_i and x_{i+1} are the lower and upper limits for the interval i and the deaths d_i are the corresponding frequencies, for $i = \alpha, \alpha + 1, \cdots, w$. The sum of the frequencies equals the number of survivors at age x_α, or

$$d_\alpha + \cdots + d_w = l_\alpha .$$

The total number of years remaining to the l_α survivors depends on the exact age at which death occurs, or on the distribution of deaths within each age interval.

Suppose that the distribution of deaths in each interval is such that, on the average, each of the d_i persons lives $a_i n_i$ years in the age interval (x_i, x_{i+1}), where a_i is a fractional number, then on the average each of the d_i persons will have lived $x_i + a_i n_i$ years, or $x_i - x_\alpha + a_i n_i$ years after age x_α, and the observed expectation of life at age x_α is obviously

$$\hat{e}_\alpha = \frac{1}{l_\alpha} \sum_{i=\alpha}^{w} (x_i - x_\alpha + a_i n_i) d_i , \quad \text{for} \quad \alpha = 0, 1, \cdots, w. \tag{53}$$

Using the relationship $d_i = l_i - l_{i+1}$, and arranging terms, we have a general formula for the observed expectation of life,

$$\begin{aligned} \hat{e}_\alpha &= a_\alpha n_\alpha + \sum_{i=\alpha+1}^{w} c_i \frac{l_i}{l_\alpha} \\ &= a_\alpha n_\alpha + \sum_{i=\alpha+1}^{w} c_i \hat{p}_{\alpha i} , \quad \text{for} \quad \alpha = 0, 1, \cdots, w, \end{aligned} \tag{54}$$

where $c_i = (1 - a_{i-1}) n_{i-1} + a_i n_i$, for $i > \alpha$. Now, if $n_i = n$, for $i = \alpha, \alpha + 1, \cdots, w$, and if the distribution of deaths in each interval is assumed to be uniform so that $a_i = \frac{1}{2}$, then $c_i = n$ and (54) reduces to

$$\hat{e}_\alpha = \frac{n}{2} + \frac{n(l_{\alpha+1} + \cdots + l_w)}{l_\alpha} \tag{55}$$

a formula often used to compute the observed expectation of life at age x_α.

Clearly, under the respective assumptions regarding the distribution of deaths in each age interval, the observed expectation of life given in formula (54) or (55) is an unbiased estimator of the corresponding true expectation of life as expressed in formula (51). On the other hand, because the mathematical expectation of the ratio of survivors, l_i to l_α, is equal to the ratio of their expectations, $l_0 p_{0i}$ to $l_0 p_{0\alpha}$, as shown in formula (41), the mathematical expectation of the observed expectation of life as given by (54) is simply

$$e_\alpha = a_\alpha n_\alpha + \sum_{i=\alpha+1}^{w} c_i p_{\alpha i}\,, \quad \text{for} \quad \alpha = 0, 1, \cdots, w. \tag{56}$$

Formula (56) will be used in developing the formula for the variance of $\hat{e}_\alpha$. As a further aid in deriving the variance of $\hat{e}_\alpha$, it is convenient to note the relationship between e_{i+1} and e_i, the expectation of life at the begiinnng of two consecutive intervals,

$$e_i - a_i n_i = [e_{i+1} + (1 - a_i)n_i]p_i\,, \quad \text{for} \quad i = 1, \cdots, w - 1. \tag{57}$$

The variance of the observed expectation of life is obtained from (54), expressing $\hat{e}_\alpha$ as a linear function of the proportions of survivors. Thus its variance is

$$\sigma^2_{\hat{e}\alpha} = \sum_{i=\alpha+1}^{w} c_i^2 \sigma^2_{\hat{p}_{\alpha i}} + 2 \sum_{i=\alpha+1}^{w-1} \sum_{j=i+1}^{w} c_i c_j \sigma_{\hat{p}_{\alpha i}, \hat{p}_{\alpha j}}\,, \quad \text{for } \alpha = 0, \cdots, w. \tag{58}$$

Substituting formula (48) in (58), we have

$$\sigma^2_{\hat{e}\alpha} = E\left(\frac{1}{l_\alpha}\right)\left[\sum_{i=\alpha+1}^{w} c_i^2 p_{\alpha i}(1 - p_{\alpha i}) + 2 \sum_{i=\alpha+1}^{w-1} \sum_{j=i+1}^{w} c_i c_j p_{\alpha j}(1 - p_{\alpha i})\right]$$

$$= E\left(\frac{1}{l_\alpha}\right)\left[\sum_{i=\alpha+1}^{w} c_i^2 p_{\alpha i} + 2 \sum_{i=\alpha+1}^{w-1} c_i \sum_{j=i+1}^{w} c_j p_{\alpha j} - \left(\sum_{i=\alpha+1}^{w} c_i p_{\alpha i}\right)^2\right]. \tag{59}$$

Using the relation $p_{\alpha j} = p_{\alpha i} p_{ij}$ and formula (56),

$$\sigma^2_{\hat{e}\alpha} = E\left(\frac{1}{l_\alpha}\right)\left[\sum_{i=\alpha+1}^{w} c_i^2 p_{\alpha i} + 2 \sum_{i=\alpha+1}^{w} c_i p_{\alpha i}(e_i - a_i n_i) - (e_\alpha - a_\alpha n_\alpha)^2\right]$$

$$= E\left(\frac{1}{l_\alpha}\right)\left[\sum_{i=\alpha+1}^{w} \{c_i(c_i - 2a_i n_i) + 2c_i e_i\} p_{\alpha i} - (e_\alpha - a_\alpha n_\alpha)^2\right]. \tag{60}$$

Since $c_i = (1 - a_{i-1})n_{i-1} + a_i n_i$, the quantity inside the braces may be rewritten as

$$\begin{aligned} c_i(c_i - 2a_i n_i) + 2c_i e_i & \\ &= [(1 - a_{i-1})^2 n_{i-1}^2 - a_i^2 n_i^2] + 2[(1 - a_{i-1})n_{i-1} + a_i n_i]e_i \\ &= [e_i + (1 - a_{i-1})n_{i-1}]^2 - [e_i - a_i n_i]^2 \\ &= [e_i + (1 - a_{i-1})n_{i-1}]^2 - [e_{i+1} + (1 - a_i)n_i]^2 p_i^2\,. \end{aligned}$$

Substitution of the last expression in (60) gives

$$\sigma^2_{\hat{e}_\alpha} = E\left(\frac{1}{l_\alpha}\right)\left[\sum_{i=\alpha+1}^{w} [\{e_i + (1 - a_{i-1})n_{i-1}\}^2 p_{\alpha i} - \{e_{i+1} + (1 - a_i)n_i\}^2 p_i^2 p_{\alpha i}] - (e_\alpha - a_\alpha n_\alpha)^2\right]. \quad (61)$$

Making the substitutions of $p_{\alpha, w+1} = 0$ and $(e_\alpha - a_\alpha n_\alpha) = [e_{\alpha+1} + (1 - a_\alpha)n_\alpha]p_\alpha$ in (61) and combining terms,

$$\sigma^2_{\hat{e}_\alpha} = E\left(\frac{1}{l_\alpha}\right)\left[\sum_{i=\alpha}^{w} [\{e_{i+1} + (1 - a_i)n_i\}^2 p_{\alpha, i+1} - \{e_{i+1} + (1 - a_i)n_i\}^2 p_i^2 p_{\alpha i}]\right].$$

Since $p_{\alpha, i+1} = p_{\alpha i}p_i$, we have the final formula for the variance of the observed expectation of life at age x_α,

$$\sigma^2_{\hat{e}_\alpha} = E\left(\frac{1}{l_\alpha}\right)\left[\sum_{i=\alpha}^{w} \{e_{i+1} + (1 - a_i)n_i\}^2 p_{\alpha i}p_i(1 - p_i)\right], \quad \text{for } \alpha = 0, 1, \cdots, w - 1. \quad (62)$$

When $E(1/l_\alpha)$ is approximated by $1/E(l_\alpha)$, $p_{\alpha i}$ is written for $E(l_i)/E(l_\alpha)$, and (43) is used for the variance of $\hat{q}_i$, formula (62) is reduced to

$$\sigma^2_{\hat{e}_\alpha} = \sum_{i=\alpha}^{w-1} p^2_{\alpha i}[e_{i+1} + (1 - a_i)n_i]^2 \sigma^2_{\hat{q}_i}, \quad \text{for } \alpha = 0, 1, \cdots, w - 1. \quad (63)$$

Thus we have proved a rather useful theorem in the study of the life table.

Theorem. If the distribution of deaths in the age interval (x_i, x_{i+1}) *is such that, on the average, each of the* d_i *individuals lives* $a_i n_i$ *years in the interval, for* $i = \alpha, \alpha + 1, \cdots, w$, *then as* l_α *approaches infinity, the probability distribution of the observed expectation of life at age* x_α *as given by* (54) *is asymptotically normal and has the mean and the variance as given by* (56) *and* (63), *respectively.*

It should be noted that (63), which is an approximation to the exact formula (62) for the variance of $\hat{e}_\alpha$ when l_α is a random variable, is identical with (62) when l_α is a given constant, such as l_0.

As a matter of practical interest, the following corollary deserves particular mention.

Corollary: If the age interval is constant, that is, if $n_i = n$, *and if the distribution of deaths in each interval is such that, on the average,* *each of the d_i individuals lives half the interval (x_i, x_{i+1}), for $i = \alpha$, $\alpha + 1, \cdots, w$, then the variance of the observed expectation of life at age x_α as given by* (55) *is*

$$\sigma^2_{\hat{e}_\alpha} = E\left(\frac{1}{l_\alpha}\right)\left[2n \sum_{i=\alpha}^{w} p_{\alpha i} e_i - \left(e_\alpha + \frac{n}{2}\right)^2\right], \quad \textit{for} \quad \alpha = 0, \cdots, w. \tag{64}$$

Proof. When $n_i = n$ and $a_i = \frac{1}{2}$, $c_i = n$ and $c_i - 2a_i n_i = 0$. From (60) we have

$$\sigma^2_{\hat{e}_\alpha} = E\left(\frac{1}{l_\alpha}\right)\left[2n \sum_{i=\alpha+1}^{w} p_{\alpha i} e_i - \left(e_\alpha - \frac{n}{2}\right)^2\right],$$

which can be rewritten as (64).

Remark 3: Although the theorem concerning the asymptotic distribution of the observed expectation of life is true for the cohort and the current life table, it is not clear why formula (63) holds also for the latter case. In the current life table, the basic random variable $\hat{q}_i$ is computed from actual mortality experience and, in general, its variance is not given by either formula (42) or formula (43). Therefore it is essential to prove (63) from the viewpoint of the current life table.[4]

The observed expectation of life, as given in (54), is a linear function of $\hat{p}_{\alpha j}$, which, in the current life table, is computed from

$$\hat{p}_{\alpha j} = \hat{p}_\alpha \hat{p}_{\alpha+1} \cdots \hat{p}_{j-1}, \quad \text{for} \quad j = \alpha + 1, \cdots, w. \tag{65}$$

Clearly, the derivatives taken at the true point $(p_\alpha, p_{\alpha+1}, \cdots, p_{j-1})$ are

$$\left[\frac{\partial}{\partial \hat{p}_i} \hat{p}_{\alpha j}\right] = p_{\alpha i} p_{i+1,j}, \quad \text{for} \quad \alpha \leqq i < j;$$

$$= 0, \qquad \text{for} \quad i \geqq j, \tag{66}$$

and hence

$$\left\{\frac{\partial}{\partial \hat{p}_i} \hat{e}_\alpha\right\} = \sum_{j=i+1}^{w} c_j p_{\alpha i} p_{i+1,j}$$

$$= p_{\alpha i}\left[c_{i+1} + \sum_{j=i+2}^{w} c_j p_{i+1,j}\right]$$

$$= p_{\alpha i}[e_{i+1} + (1 - a_i)n_i]. \tag{67}$$

[4]Using a different approach [12], professor E. B. Wilson derived the following formula for the variance of $\hat{e}_\alpha$,

$$\sigma^2_{\hat{e}_\alpha} = \frac{1}{l_\alpha^2} \sum_{i=\alpha}^{w-1} l_i^2 [e_{i+1} + a_i n_i]^2 \sigma^2_{\hat{q}_i},$$

which is in error in that the quantity $a_i n_i$ should be replaced by $(1 - a_i)n_i$.

Because of (66), the derivative (67) vanishes when $i = w$. Since it has been shown in Lemma 3 that the covariance between proportions of survivors of two non-overlapping age intervals is zero, the variance of the observed expectation of life may be computed from the following:

$$\sigma^2_{\hat{e}_\alpha} = \sum_{i=\alpha}^{w-1} \left[\frac{\partial}{\partial \hat{p}_i} \hat{e}_\alpha \right]^2 \sigma^2_{\hat{p}_i} . \tag{68}$$

Substitution of (67) in (68) gives formula (63).

When the distribution of deaths within each age interval is assumed to be uniform, $a_i = \frac{1}{2}$, and (63) becomes

$$\sigma^2_{\hat{e}_\alpha} = \sum_{i=\alpha}^{w-1} p^2_{\alpha i} \left[e_{i+1} + \frac{n_i}{2} \right]^2 \sigma^2_{\hat{p}_i} , \quad \text{for} \quad \alpha = 0, 1, \cdots, w. \tag{69}$$

REFERENCES

[1] Chiang, C. L. [1954]. Competition and other interactions between species. *Statistics and Mathematics in Biology,* 197–215, Iowa State University Press, Ames, Iowa.

[2] Chiang, C. L. [1957]. An application of stochastic processes to experimental studies of flour beetles. *Biometrics 13,* 79–97.

[3] Chiang, C. L. [1958]. An application of stochastic processes to the life table and standard error of age-adjusted rates. (abstract) *Biometrics 14,* 133–4.

[4] Chiang, C. L. [unpublished]. A practical example where the random variables are uncorrelated but not independently distributed.

[5] Dublin, L. I., Lotka, A. J., and Spiegelman, M. [1949]. *Length of Life.* New York: Ronald Press Co.

[6] Feller, W. [1950]. *An Introduction to Probability Theory and Its Applications.* New York: John Wiley & Sons, Inc.

[7] Greenwood, M. [1926]. A report on the natural duration of cancer. *Report on Public Health and Medical Subjects, No. 33,* 1–26, His Majesty's Stationary Office.

[8] Greville, T. N. E. [1943]. Short methods of constructing abridged life table. *Rec. Amer. Inst. Actu. 32,* 29–42.

[9] Halley, E. [1693]. An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslau. *Philos. Trans. 17,* 596–610, Royal Society of London.

[10] Irwin, J. O. [1949]. The standard error of an estimate of expectational life. *J. Hyg. 47,* 188–9.

[11] Reed, L. J., and Merrell, M. [1939]. A short method for constructing an abridged life table. *Amer. J. Hyg. 30,* 33–62.

[12] Wilson, E. B. [1938]. The standard deviation of sampling for life expectancy. *J. Amer. Stat. Assoc. 33,* 705–8.