*Measurement of duration as a characteristic of a cohort is subject to biases unless certain pitfalls discussed in this paper are avoided. The authors discuss aspects of this problem and indicate its significance for health studies.*

# ON RELATIONSHIPS BETWEEN LONGITUDINAL CHARACTERISTICS AND CROSS-SECTIONAL DATA

*Jane A. Menken, M.S., and Mindel C. Sheps, M.D., M.P.H., F.A.P.H.A.*

## Introduction

THE duration of a condition is a frequently studied longitudinal characteristic of a cohort. It is an aspect of experience that is of great analytic importance in medicine, demography, and doubtless many other fields. Certain problems in the measurement of duration are common to a wide variety of circumstances. In particular, in this paper, it will be shown that unless certain nonobvious pitfalls are avoided, duration measurement is subject to serious biases.

A wide variety of duration variables has been investigated. For example, in demography and public health, duration of life has been studied extensively. In clinical situations, interest has focused on the duration of the preclinical stage of a disease,[1] or on survival time, i.e., on duration of life after diagnosis of a condition.[2] The duration of hospital stay is a variable of great concern in meeting medical care needs.[3] Other studies have examined the duration between recurrences of an event or condition which can occur to a given individual more than once. Thus, the lengths of the intervals between episodes, or recurrences, of an illness (i.e., the durations of remissions) have been investigated.[4] Marriage can occur repeatedly to a given individual, and its duration according to marital order (first, second, third, marriage for husband or wife) is of interest.[5] In demography, birth intervals, i.e., the durations of intervals between marriage and first birth, and between successive births, are studied[6,7] as one aspect of natality analysis.

In exploring some of the problems inherent in the measurement of duration, we will assume that the duration variable has a distribution which does not change with time, so that the distribution is the same for all incidence groups, i.e., for all cohorts. When we cannot assume identical distributions, the difficulties that we will point out become more severe. Hence, making this assumption merely facilitates the presentation.

In some cases, investigators have attempted to measure duration from prevalence cases. Therefore, in the next section, we consider what is known as the stable disease model and review the relationship, well known in probability theory but generally unfamiliar to epidemiologists, between the distribution of duration in an incidence group and the distribution of duration among cases prevalent at a specified time. Implications of these results are briefly examined. Even when the assumption of a stable disease model is tenable, it is usually necessary to measure duration by following an incidence group. Particularly when a condition occurs repeatedly, the duration of time for which the cohort is followed is an important, but often unrecognized, determinant of the distribution of the observed duration of the condition. In the third section, the effects of this factor will be examined theoretically and numerically, using results from a computer simulation model for reproduction.

## Duration of Prevalence Cases in a Stable Disease

The stable disease model assumes[8] that (1) the distribution of duration is the same for all incidence cohorts; (2) incidence is constant over time, i.e., the same number of cases begin in each time unit; (3) the disease has a finite maximum duration.

These assumptions lead to constant prevalence. Although it would perhaps be more accurate to refer to this situation as a stationary disease model, since the disease population reaches and remains at a constant size, we will use the customary terminology.

A simple example of a stable disease is given in Figure 1. One-third of the cases last one day; another third last two days and the remainder, three days. In other words, the probability distribution function (p.d.f.) of duration in each incidence cohort is given by

$$f(x) = \begin{cases} 1/3 & x=1,\ 2,\ 3 \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Mean duration, $\mu_x$, is two days and variance, $\sigma_x^2$, equals $2/3$. When incidence is equal to three per day, the illustrated disease pattern follows. Since duration is taken here as a discrete variable, it is necessary to be quite precise in defining a prevalent case. Cases newly incident on a specific day are not counted as being prevalent on that day. Letting $x_m$ be the maximum duration described under assumption (3), we see that prevalence becomes constant on Day $x_m+1$ or, in our example, on Day Four. Then, for example, on Day Six, prevalence is equal to six cases (numbered 1 to 6 in Figure 1) and incidence is equal to three (A, B, C), so that the well-known relationship that prevalence is equal to incidence multiplied by mean duration indeed holds.

Information about the detailed *distribution* of duration cannot come simply from prevalence and incidence data, but requires actual duration measurements. Using a cross-sectional approach,

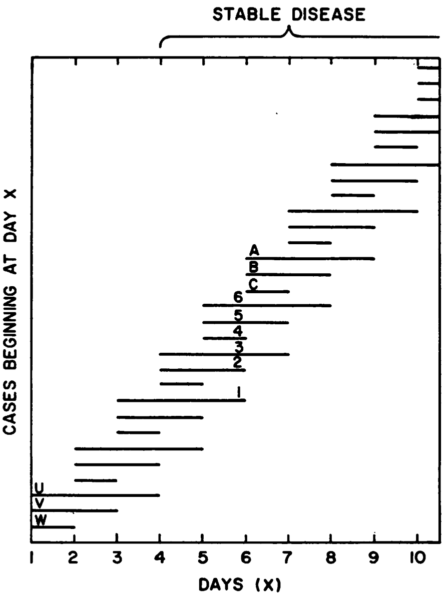


**Figure 1—Simple example of the stable disease model. Each line in the figure represents a case. The line runs from the day of incidence through the last day of the illness**

the duration of prevalence cases, which differs from that of incidence cases, could be measured at a given point in time, t.

Two related variables may be studied among prevalence cases: (1) the duration, U, of the open interval between the onset of the disease and t, and (2) the total duration, T, of cases which are prevalent at time t and are then followed to their conclusion. U has been measured with the idea that the length of the open interval for prevalence cases represents half the total duration of incidence cases. Unfortunately, this is not the case. Turning again to Figure 1, we find that if the open interval is measured at Day Six, the open interval is three days for Case 1; two days for Cases 2 and 3, and one day for Cases 4, 5, and 6, i.e., the p.d.f. is

$$h(u) = \begin{cases} 3/6 & u=1 \\ 2/6 & u=2 \\ 1/6 & u=3 \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

with mean $\mu\# = 1\frac{2}{3}$ days. Obviously $\mu\#$ is greater than $\frac{\mu_x}{2}$.

T, the total duration of the prevalence cases, is

| Case No. | Duration in days |
|---|---|
| 1 | 3 |
| 2 | 2 |
| 3 | 3 |
| 4 | 1 |
| 5 | 2 |
| 6 | 3. |

Therefore, the p.d.f. of T is:

$$g(x) = \begin{cases} 1/6 & x=1 \\ 2/6 & x=2 \\ 3/6 & x=3 \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

and the mean duration, $\mu^*$, of prevalence cases is greater than $\mu_x$, specifically, $\mu^* = 2\frac{1}{3}$ days.

Explanations for these related phenomena can be reached through the use of renewal theory. The stable disease model, because of its constant incidence and unchanging distribution of duration, satisfies the definition of a renewal process.[9-11] Results from renewal theory can be applied to clarify the relations between the distributions f(x), h(u), and g(x) at any time $t > x_m$. It can be shown (and proofs are given in Appendix 1), that when the distribution of duration of incidence cases is given by f(x), the p.d.f. of U, the length of the open interval, is

$$h(u) = \frac{\sum_{x=u}^{x_m} f(x)}{\mu_x} \tag{4}$$

with mean

$$\mu\# = \frac{\mu_x}{2} + \frac{\sigma^2}{2\mu_x} + \frac{1}{2}. \tag{5}$$

The p.d.f. of T, the duration of prevalence cases, is given by

$$g(x) = \frac{x f(x)}{\mu_x} \qquad x = 0, 1, 2 \tag{6}$$

with mean

$$\mu^* = \sum_{x=0}^{\infty} \frac{x^2 f(x)}{\mu_x} = \frac{\sigma_x^2 + \mu_x^2}{\mu_x}$$

$$= \mu_x + \frac{\sigma_x^2}{\mu_x}. \tag{7}$$

Comparison of (5) and (7) shows that

$$\mu\# = \frac{1}{2}\,\mu^* + \frac{1}{2}, \tag{8}$$

where the term ½ comes from the fact that we are dealing with a discrete distribution.* From (6) we see that prevalence cases can be considered a "length-biased" sample because the probability of a case being prevalent at a given time is proportional to its duration times its probability. Therefore, the mean duration of prevalence cases is greater than that of incidence cases and the mean open interval is greater than half the mean duration of incidence cases. Although Zelen and Feinleib[1] have made

* Results comparable to (4)-(7) for continuous distributions can be found in the books by Cox and Feller, cited in References 9-11.

ingenious use of these and other relationships derived from renewal theory to estimate related duration variables in a two-stage disease model, these results generally have been difficult to apply. In addition, the assumption of a constant number of incidence cases is rarely valid. Consequently, it is usually necessary to follow an incidence cohort in order to obtain accurate measures of duration parameters. The remainder of this paper considers the effects on duration measurement of the specific point in time at which the measurements on the cohort are made.

## Duration Measured in a Cohort

Intuitively, it is obvious that, unless a cohort is followed until every member has reached the end of the condition, the observed distribution of duration will be incomplete. The higher values of duration, X, will not be observed and the mean, $\overline{X}$, of observed completed intervals will be less than $\mu_x$, the true mean of the distribution. Referring again to Figure 1, consider the cohort starting on Day 1, labeled U, V, and W. If duration is measured on Day 3, only V and W will have completed the interval. Their mean duration is one and a half days. Only if duration is measured on Day 4, or later, after the intervals for all members of the cohort have ended, i.e., at some $t > x_m$, will the true mean of two days be observed.

When duration of a condition can be long, this problem is especially important. Investigators studying birth intervals have encountered some of its facets in attempting to analyze data and in developing probability models for reproduction.[12,13] Birth intervals can be thought of as the result of a number of underlying factors that lead to a distribution, denoted by $f_i(x)$, of the duration of the interval between the $\overline{(i-1)}$th and the ith live birth. For our current purpose (to examine the effect of the time at which duration variables are measured), birth intervals are used only as an example.

Consider the interval, Y, between marriage and the fourth birth in a marriage cohort. This interval has four segments, as illustrated in Figure 2. Let us denote by

$X_1$ the length of the interval from marriage to first birth

$X_i$ the length of the interval from $(i-1)$th to ith birth, $i \geq 2$

$f_i(x)$ the p.d.f. of $X_i$, $\geq 1$

Y the length of the interval between marriage and fourth birth, where

$$Y = \sum_{i=1}^{4} X_i$$

and $h_4(y)$ the p.d.f. of Y.

Then the mean, or expected value of Y, is

$$EY = EX_1 + EX_2 + EX_3 + EX_4. \qquad (9)$$

If we make the simplifying assumption that the $X_i$ are independent, then Var Y is the sum of the variances of the $X_i$, and $h_4(y)$, which is given in Appendix 2, is the convolution of the $f_i(x)$. It is important to stress that $h_4(y)$ is the distribution which would result if all women were followed to their fourth birth. The p.d.f., when women are followed for a shorter specified length of time, is also derived in Appendix 2. This distribution is difficult to study analytically. A Monte Carlo simulation program developed by Ridley and Sheps, REPSIM A,[14] was used to obtain numerical results for the mean time to the fourth birth, when a marital cohort was followed for 5, 10, 15, and

**Figure 2—Interval from marriage to the fourth live birth (i.e., to the fourth "episode")**

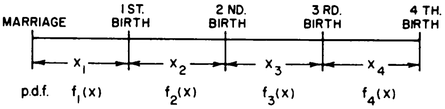


Table 1—Mean interval to the fourth birth, by duration of a marriage, in a marital cohort of 1,000[a] women

| Marital duration | 5 yr | 10 yr | 15 yr | 20 yr |
|---|---|---|---|---|
| 1. No. of women of parity 4+ | 5 | 527 | 986 | 1,000 |
| 2. Mean interval to 4th birth (in mo) | 50.8 | 98.9 | 116.9 | 118.6 |

a. Expected value = 118.0.

20 years, and the following assumptions concerning the $X_i$ were made:

1) $X_1$ has a distribution $f_1(x)$ with mean 16.3 months and variance 32, and
2) the $X_i$, $i \geq 2$, have identical distributions, denoted $f(x)$, with mean 33.9 months and variance 214.

Then Y has distribution $h_4(y)$ with mean 118.0 months and variance 674.

In Table 1, simulated results are presented which show the effect of measuring intervals before all members of a cohort have completed that interval. Line 1 shows the number of women reaching parity 4. After 5 years of marriage, only 5 women have had four or more births. This number increases to 1,000 as marital duration increases to 20 years. The mean observed value of Y is only 50.8 months at 5 years, a value far below its theoretical expectation of 118 months. This observed mean increases with marital duration to 118.6 months at 20 years. The difference from the true expected mean (118.0 months) is due to the fact that we are dealing with simulated results.

The results for the individual intervals, the $X_i$, are also affected, as shown in Table 2. As duration of marriage increases, the number of women of parity 3 or more (line 1) also increases. The mean of the interval between the second and third births is given on line 2. This quantity, $\bar{X}_3$, increases with marital duration as does $\bar{X}_4$, shown on line 5. Comparison of $\bar{X}_3$ and $\bar{X}_4$ shows that, at every marital duration where not all women have had at least four births, $\bar{X}_3$ is greater than $\bar{X}_4$. At 20 years, these values are the same, within sampling error due to simulation. This results from the fact that, at shorter marital durations, only women with relatively short intervals can have a large number of births. In fact, if we look at the mean of $\bar{X}_3$ for women who have had at least four births (line 7), we find it is shorter, at every marital duration except 20 years, than $\bar{X}_3$, but longer than $\bar{X}_4$. Sheps, et al.,[15] have explored this problem in much greater detail, and have given further results related to the open interval since the most recent birth.

## Conclusions

Although our discussion has used birth intervals as an example, it is equally applicable to the measurement of variables, such as the duration of the interval between successive hospitalizations for mental illness, duration of successive marriages, duration of residence, and the like.

To summarize, let us consider single episode duration variables first. Two points are clear. Prevalence cases are a biased sample. In the stable disease model, the cases with longer duration are more likely to be included in a sample of prevalence cases. In a cohort, the bias, unless all cases are followed to their conclusion, is in the opposite direction.

If a condition occurs repeatedly, the observed duration of any given interval or episode depends upon the length of follow-up, with the mean duration increasing until all members of a cohort have completed that interval. Also, only

Table 2—Mean interval between second and third birth and between third and fourth birth, by duration of marriage, in a marital cohort of 1,000 women[b]

| Marital duration | 5 yr | 10 yr | 15 yr | 20 yr |
|---|---|---|---|---|
| 1. No. of women of parity 3+ | 124 | 940 | 1,000 | 1,000 |
| Interval between second and third birth, $X_3$, for women of parity 3+ | | | | |
| 2. Mean, $\overline{X}_3$ | 19.4 | 34.0 | 34.7 | 34.3 |
| 3. Variance | 41.0 | 185.3 | 204.9 | 212.6 |
| 4. No. of women of parity 4+ | 5 | 527 | 986 | 1,000 |
| Interval between third and fourth birth, $X_4$, for women of parity 4+ | | | | |
| 5. Mean, $\overline{X}_4$ | 12.6 | 27.0 | 32.7 | 34.4 |
| 6. Variance | 3.8 | 112.6 | 192.2 | 223.0 |
| Interval between second and third birth for women of parity 4+ | | | | |
| 7. Mean | 13.4 | 29.3 | 34.3 | 34.3 |
| 8. Variance | 2.2 | 133.9 | 193.3 | 212.6 |

[b] The distribution of the intervals between successive births are independent and have theoretical means and variances:

$EX_1 = 16.3$ months
$\text{Var } X_1 = 32$

$EX_i = 33.9$
$\text{Var } X_i = 214$   $i \geq 2$

individuals with short durations for each episode can have a large number of episodes in a given time period.

These findings show that the effect of time of measurement can produce a spurious difference, particularly for a recurring condition, even when the duration of each episode has the same theoretical distribution; or it can perhaps mask true differences. It should also be emphasized that great care must be taken in defining the duration variable of interest. Variables such as the duration of a current episode or of the most recent episode, even for an identifiable cohort, are mixtures of durations of first, second, and succeeding episodes. If such variables are used, their structure must be taken into consideration in any analysis or when making comparisons between groups.

## REFERENCES

1. Zelen, M., and Feinleib, M. On the Theory of Screening for Chronic Diseases. Biometrika (in press).
2. Merrell, M., and Shulman, L. E. Determination of Prognosis in Chronic Disease, Illustrated by Systemic Lupus Erythematosis. J. Chronic Dis. 1:12-32, 1955.
3. Rosenfeld, L. S.; Goldmann, F.; and Kaprio, L. A. Reasons for Prolonged Hospital Stay. Ibid. 6:141-152, 1957.
4. Fairbanks, V.; Shanbron, E.; Steinfeld, J.; and Beutler, E. Prolonged Remissions in Acute Myelocytic Leukemia in Adults. J.A.M.A. 204:574-579, 1968.
5. Divorce Statistics Analysis: United States, 1964 and 1965. PHS Publ. No. 1000, Ser. 21-17. Washington, D. C.: Gov. Ptg. Office (Oct.), 1969.
6. Henry, L. Fécondité et Famille—Modèles Mathématiques II: Applications numériques. Population 16:261-282, 1961.
7. Sheps, M. C. An Analysis of Reproductive Patterns in an American Isolate. Population Studies 19:65-80, 1965.

8. Feinleib, M. The Stable Disease Model (abstract). Biometrics 23:1299, 1967.
9. Feller, W. An Introduction to Probability Theory and Its Applications, Vol. I (2nd ed.). New York: Wiley, 1957.
10. Cox, D. R. Renewal Theory. New York: Barnes and Noble, 1962.
11. Feller, W. An Introduction to Probability Theory and Its Applications, Vol. II. New York: Wiley, 1966.
12. Henry, L. Fécondité et Famille—Modèles Mathématiques II. Population 16:27-48, 1961.
13. Potter, R. G.; New, M. L.; Wyon, J. B.; and Gordon, J. E. "Applications of Field Studies to Research on the Physiology of Human Reproduction: Lactation and Its Effects Upon Birth Intervals in Eleven Punjab Villages, India." In: M. C. Sheps and J. C. Ridley (eds.). Public Health and Population Change. Pittsburgh, Pa.: University of Pittsburgh Press, 1965.
14. Ridley, J. C., and Sheps, M. C. An Analytic Simulation Model of Human Reproduction with Demographic and Biological Components. Population Studies 19:297-310, 1966.
15. Sheps, M. C.; Menken, J. A.; Ridley, J. C.; and Lingner, J. L. The Truncation Effect in the Analysis of Closed and Open Birth Interval Data. J. Am. Statist. A. 65, 1970.

## APPENDIX 1

### Distribution of Prevalence Cases in a Stable Disease Model

*Notation and Definitions*

$I(t)$ number of cases with incidence at t. $I(t)$ is assumed to be constant. i.e., $I(t)=I$, all $t>0$. These cases are not counted as prevalent at time t

$X(t)$ variable duration of cases in cohort with incidence at t

$f(x)$ probability distribution function (p.d.f.) of $X(t)$, which is assumed to be the same for all t

$\mu_x$ mean duration of cases in an incidence cohort, i.e.,

$$\mu_x = \sum_{x=1}^{\infty} xf(x)$$

$x_m$ maximum value of $X(t)$, i.e., $f(x)=0$ for all $x>x_m$

$\sigma_x^2$ variance of duration, i.e.,

$$\sigma_x^2 = \sum_{x=1}^{\infty} (x-\mu_x)^2 f(x)$$

$P(t)$ number of cases prevalent at t

$U|t$ variable duration of open interval foı cases prevalent at t

$h(u|t)$ p.d.f. of $U|t$

$T|t$ variable total duration of cases prevalent at t

$g(x|t)$ p.d.f. of $T|t$.

*Prevalence at Time t*

We will first derive an expression for $P(t)$ and show that, as $t \to \infty$, $P(t)$ approaches a constant.

Consider Figure 3, the open interval $U|t$, for a case with incidence at t-7. Obviously, its duration is at least 7 time units. In general, a case with incidence at $t-x$ is prevalent at t if its duration is $x, x+1, x+2, \ldots$ Hence, the number of cases with incidence at $t-x$ which are prevalent at t is given by

$$I(t-x) \sum_{y=x}^{\infty} f(y). \qquad (1.1)$$

By summing (1.1) over all possible values of x, we find

$$P(t) = \sum_{x=1}^{t-1} [I(t-x) \sum_{y=x}^{\infty} f(y)]\,. \qquad (1.2)$$

Changing the order of summation,

$$P(t) = \sum_{x=1}^{\infty} [f(x) \sum_{y=1}^{c} I(t-y)] \qquad (1.3)$$

$$= I \sum_{x=1}^{\infty} cf(x) \qquad (1.4)$$

$$c=\min(t-1,x),$$

since $I(t-y)=I$ for all $(t-y)>0$. As $t \to \infty$, $c \to x$, so that

$$P(t) \to P = I \sum_{x=1}^{\infty} xf(x)$$

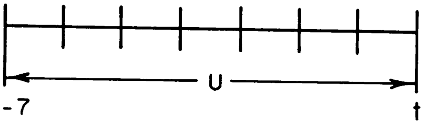
$$= I\mu_x. \qquad (1.5)$$

Then

$$I = P/\mu_x. \qquad (1.6)$$

(1.6) is one version of the well-known stable disease relationship between incidence, prevalence, and mean duration. Since we have assumed $f(x)=0$ for $x>x_m$, (1.5) and (1.6) hold for all $t>x_m$.

**Figure 3—Open interval, U|t, for a case with incidence at t−7**



### *P.d.f. of Open Interval for Cases Prevalent at t*

The number of cases with open interval $U|t=x$ is given by (1.1). Letting $F(x)=\sum_{y=x}^{\infty} f(y)$, we find from (1.1) and (1.2) that the p.d.f. of $U|t$ is

$$h(u|t)=\frac{I(t-u)F(u)}{P(t)}=\frac{F(u)}{\sum_{u=1}^{t-1} F(u)}. \qquad (1.7)$$

Expanding the denominator of (1.7), we see that as t → ∞,

$$\sum_{u=1}^{t-1} F(u) \to f(1)+2f(2)+\ldots+nf(n)+\ldots$$

$$=\sum_{u=1}^{\infty} uf(u)$$

$$=\mu_x. \qquad (1.8)$$

Therefore, as t → ∞

$$h(u|t) \to \frac{F(u)}{\mu_x}. \qquad (1.9)$$

The mean open interval, as t → ∞, becomes

$$\mu\# = \frac{\sum_{u=1}^{\infty} uF(u)}{\mu_x} = \frac{\sum_{u=1}^{\infty} f(u)[\sum_{x=1}^{u} x]}{\mu_x}$$

$$= \frac{\sum_{u=1}^{\infty} u(u+1)\, f(u)}{2\mu_x}$$

$$= \frac{\sum_{u=1}^{\infty} u^2 f(u) - \mu_x^2 + \mu_x^2 + \sum_{u=1}^{\infty} uf(u)}{2\mu_x}$$

$$= \frac{\sigma_x^2 + \mu_x^2 + \mu_x}{2\mu_x}$$

$$= \frac{\mu_x}{2} + \frac{\sigma_x^2}{2\mu_x} + \frac{1}{2}. \qquad (1.10)$$

### *P.d.f. of Cases Prevalent at t*

Note that, in (1.3), the expression within brackets is equal to the number of prevalence cases, at t, with total duration x. Therefore

$$g(x|t)=\frac{f(x)\sum_{y=1}^{c} I(t-y)}{P(t)}$$

$$=\frac{I\, cf(x)}{P(t)}$$

$$=\frac{P}{P(t)}\;\frac{cf(x)}{\mu_x}. \qquad (1.11)$$

$$c=\min\,(t-1,\ x)$$

The last equality follows by substitution from (1.6).

Finally, as t → ∞, c → x and P(t) → P so that

$$g(x|t) \to g(x) = \frac{xf(x)}{\mu_x}. \qquad (1.12)$$

Then, as t → ∞, the probability of a prevalence case having duration x is constant for all t and proportional to the duration x times the probability, f(x), that an incidence case has duration x. The mean duration of prevalence cases is

$$\overset{*}{\mu_x} = \sum_{x=1}^{\infty} xg(x) = \frac{1}{\mu_x}\Sigma x^2 f(x) = \frac{1}{\mu_x} E(X^2)$$

$$= \mu_x + \frac{\sigma_x^2}{\mu_x}. \qquad (1.13)$$

Again, since $f(x)=0$ for $x>x_m$, (1.8)-(1.10) and (1.12)-(1.13) hold for all $t>x_m$.

## APPENDIX 2

### Distribution of Duration of Intervals in a Cohort

#### *Assumptions*

We will derive results for the general case where the $X_i$ are lengths of intervals between successive events in a process and are mutually independent.

#### *Notation and Definitions*

$X_1$ length of interval from start of process to first event

$X_i$ length of interval following $(\underline{i-1})$th event to $\underline{i}$th event

$f_i(x)$ p.d.f. of length of the $\underline{i}$th interval

$x'_i$ minimum value of $X_i$, i.e., $\sum_{y=1}^{x'-1} f_i(y)=0$ and $f_i(x'_i)>0$

$Y$ length of interval from start of process to fourth event, i.e.,

$$Y=\sum_{i=1}^{4} X_i$$

$h_i(x)$ p.d.f. of length of interval from start of process to $\underline{i}$th event, i.e., $h_1(x)=f_1(x)$ is the p.d.f. of $X_1$

$h_i(x)$ is the p.d.f. of $\sum_{j=1}^{i} X_j$

$$H_i(x) = \sum_{y=1}^{x} h_i\ (x) = \Pr\left[\ \sum_{j=1}^{i} X_j \ \le x\ \right]$$

$t$ the time at which the measurement of interval duration is made

$t_i$ the minimum t such that $h_i(t)>0$, i.e., $t_i=\sum_{j=1}^{i} x'_j$

$X_i|t$ ith interval when measurement is taken at time t (given that the ith event has occurred)

$Y|t$ length of interval from start of process to fourth event when measurement is taken at time t (given that the fourth event has occurred)

$t_m$ the time when all members of the cohort have experienced the fourth event($t_m \leq \sum_{i=1}^{4} x_{i,m}$ where $x_{i,m}$ is the maximum value of $X_i$)

Expected values will be indicated by E preceding the variable, e.g., $EX_i$.

*Distribution of Interval to ith Event*

The distribution of $\sum_{j=1}^{i} X_j$, the time to the ith event, is the convolution of the $f_j(x)$, $j=1 \ldots i$. Then

$$h_1(x)=f_1(x)$$

$$h_2(x)=\sum_{z=1}^{x} h_1(x-z)f_2(z)$$

$$h_i(x)=\sum_{z=1}^{x} h_{i-1}(x-z)f_i(z). \qquad (2.1)$$

Specifically, the p.d.f. of $Y=\sum_{j=1}^{4} X_j$ is

$$h_4(x)=\sum_{z=1}^{x} h_3(x-z)f_4(z)\,. \qquad (2.2)$$

*Distribution of Interval to Fourth Event When Measurement Is Made at t*

The distribution of $Y|t$ agrees with that of Y only when t is sufficiently large so that all members of a cohort have experienced the fourth event, i.e. when $t \geq t_m$. Otherwise the p.d.f. of $Y|t$ is

$$h(y|t)=\frac{h_4(y)}{\sum_{z=1}^{t} h_4(z)} \qquad y\leq t,\ t<t_m \qquad (2.3)$$

with expected value

$$EY|t=\frac{\sum_{y=1}^{t} yh_4(y)}{\sum_{y=1}^{t} h_4(y)}\,. \qquad t<t_m \qquad (2.4)$$

Obviously, $EY|t < EY$ for $t<t_m$. The extent of the bias varies depending upon the specific distribution $h_4(y)$.

*Distributions of Intervals Between Events*

The behavior of $X|t$ is analogous to that of Y. For $X_i$, $i\geq 2$, the p.d.f. of $X_i|t$ is

$$f_i(x|t)=\frac{f_i(x)H_{i-1}(t-x)}{H_i(t)}\,. \qquad (2.5)$$

$$x\leq t-t_{i-1}$$

Again, $f_i(x|t)$ agrees with $f_i(x)$ only if t is sufficiently large so that all members of a cohort have the ith event with probability 1, i.e., so that $H_i(t)=1$ and $f_i(x)=0$ for all $x>t-t_{i-1}$. Otherwise, $t_i-t_{i-1}\leq x_i\leq t-t_{i-1}$.

If $f_i(x)>0$ for $x>t-t_{i-1}$, $H_i(t)<1$ and for the defined (relatively large) values of x, $f_i(x)$ is not included in the numerator of (2.5). Hence $EX_i|t < EX_i$. On the other hand, if $f_i(x)=0$ for all $x>t-t_{i-1}$ and $H_i(t)$ does not equal 1, $H_{i-1}(t-x)<1$ at least for x greater than some value $\eta$. (Otherwise, if $H_i(t-x)=1$ for all x, then $H_i(t)=1$ and $f_i(t|x)=f_i(x)$.) Consider the distribution of $H_i(t)$:

$$H_i(t)=\sum_{x=1}^{t} f_i(x)\ H_{i-1}(t-x). \qquad (2.6)$$

It follows from (2.6) that, when $H_i(t-x)<1$ for $x>\eta$, the weight given in (2.5) to $f_i(x)$ decreases with increasing x, and $EX_i|t<EX_i$. Also, if for some i, $EX_i>EX_{i-1}$, $EX|t$ is not necessarily greater than $EX_{i-1}|t$.

With increasing t, $H_i(t)$ is nondecreasing and tends to unity. $EX_i|t$ then tends to increase until it reaches its maximum value, $EX_i$.

In the case where all $f_i(x)=f(x)$, let j be the lowest index such that $EX_j|t<EX$. Then for $i\geq j$, the relative weight given to f(x) for large x in (2.5) decreases with increasing i, and $EX_i|t$ decreases with increasing $i\geq j$.

Mrs. Menken is a Research Associate at the Office of Population Research, Princeton University (5 Ivy Lane), Princeton, N. J. 08540, and Dr. Sheps is Professor of Biostatistics, University of North Carolina School of Public Health, Chapel Hill 27515

This paper was presented before the Statistics Section of the American Public Health Association at the Ninety-Seventh Annual Meeting in Philadelphia, Pa., November 11, 1969.