THE COMPARISON OF LIFE TABLE AND MARKOV CHAIN TECHNIQUES FOR FOLLOW-UP STUDIES

by

John R. Schoenfelder, Richard H. Shachtman
and Gordon J. Johnston

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1299
July 1980

THE COMPARISON OF LIFE TABLE AND MARKOV CHAIN TECHNIQUES FOR FOLLOW-UP STUDIES

by

John R. Schoenfelder, Richard H. Shachtman and
Gordon J. Johnston

ABSTRACT

We examine Markov chain versus life table techniques for comparing cohorts (homogeneous groups) with respect to time to occurrence of specified events. Key results are (1) a theorem showing that, when the underlying process is a stationary Markov chain (MC), the MC estimator of time to occurrence is equivalent to the life table estimator and (2) simulation statistics suggesting that the MC estimator has a smaller standard error. We supplement these features with a discussion of further advantages associated with using a MC modeling approach.

Key Words: Markov Chain, Life Table, Follow-up Studies, Time to Absorption (Occurrence) Distributions

Richard H. Shachtman is Professor, Department of Biostatistics, Curriculum in Operations Research and Systems Analysis, University of North Carolina, Chapel Hill, NC 27514.

John R. Schoenfelder is Visiting Research Assistant Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27514.

Gordon J. Johnston is Systems Engineer, Texas Instruments, Lewisville, TX 75067.

Research was supported in part by National Institute of Child Health and Human Development Grant HD 10367.

THE COMPARISON OF LIFE TABLE AND MARKOV CHAIN TECHNIQUES FOR FOLLOW-UP STUDIES

by

John R. Schoenfelder, Richard H. Shachtman and

Gordon J. Johnston

1. INTRODUCTION

Biological and epidemiological cohort studies often investigate the length of time between the occurrences of two specified events. Usually these events are the initiation of observation and the development of a chronic condition whose incidence is being scrutinized. The probability distribution of the resulting interval length represents the time-to-appearance of the condition. For example, for a study of the incidence of lung cancer in adults, observations might initiate at the 21st birthday of each subject and the condition attribute recorded would be the date of the first diagnosis of lung cancer. The resulting distribution of time-to-appearance of lung cancer will then reflect the propensity of members of the population represented by the study cohort to contract the disease.

A traditional approach to estimating this probability distribution is to construct a current life table, Chiang (1968), with the appearance of the condition, e.g., lung cancer, representing the death state. Such a life table (LT) represents the experience, with respect to development of the condition, of the study cohort during the period of observation and uses that representation to project the experience of the target population.

An alternative methodology for estimating the time to appearance probability distribution is proposed by Shachtman, Schoenfelder and Hogue (1978) in a study of the effect of induced abortion on subsequent pregnancy outcome. Their approach requires that the stochastic nature of the process being investigated be described by a stationary (time homogeneous) Markov Chain (MC). When this is the case, Shachtman et al. (1978) show how functions of the MC transition probability matrix may be used to compute the probability distribution of the length of time between the appearance of any two events, i.e., the length of time between visits to any two states in the MC state space.

In this paper we formally compare these two methodologies. In particular, we show that when the underlying process is a stationary MC, the MC estimator is equivalent to the LT estimator in the sense that both quantities estimate the same theoretical construct. The MC estimator, however, makes use of the Markov property whereas the LT estimator does not; hence, as will be evidenced later, the former makes more efficient use of available data.

In addition, we present several additional advantages which favor MC technology over its LT counterpart. For example, the MC approach may be used to compute the probability distribution of the length of time between visits to any two states of the MC. A LT, however, produces only the distribution function of the length of time between the initial state and (the state representing) the appearance of the condition. Thus one MC may be used to investigate several research questions, whereas each individual research question requires a separate LT. Furthermore, as seen in Shachtman, Schoenfelder and Hogue (1980), the MC formulation provides a straightforward methodology

for incorporating intervening variables into the analysis; no analogue is found in the LT approach.

Before discussing comparisons between MC and LT methodologies, we present the mathematical basis for both techniques. In this development we shall, for convenience and notational ease, assume that the condition being studied, once contracted, may not go into remission - thus it is represented by an absorbing state in the Markov chain. Since the purpose of this work is to estimate the probability distribution of time until appearance of the condition, this assumption does not prevent the methodologies from being used in a study of conditions which may go into remission. In such situations the condition state in the MC is artificially made absorbing for the purpose of analysis. The resulting probability distributions are called time to absorption (TTA) distributions.

2. TIME TO ABSORPTION DISTRIBUTIONS

We begin by developing a mathematical representation of the theoretical TTA distribution which we desire to study. We then define and construct two specific estimators of the TTA distribution - the LT estimator and the MC estimator. In the next section we formally compare these estimators.

We consider a discrete-time stochastic process $(X_t : t = 0, 1, 2, \ldots)$ which, at any given time, may occupy one of finitely many states $1, 2, \ldots, K$. Assume that r of these states are absorbing and with no loss of generality further assume that $r = 1$ and that state K is the absorbing state. (In our introductory example, entrance into state K would represent contracting lung cancer.) In order to obtain both LT and MC estimators, we assume

that the data available for our analysis is a set of statistically independent sample paths generated by this process.

Let the set of probabilities $\{a_J(0) : J = 1,2,\ldots,K;$ $\sum_{J=1}^{K} a_J(0) = 1\}$, where SUM represents the ordinary arithmetic sum, constitute the initial distribution over these states, i.e., $a_J(0) = P(X_0 = J)$. Finally, let T be the random variable representing the time (measured from 0) required for the process to first enter the absorbing state. Then the time to absorption probability distribution which we desire to estimate is the cumulative distribution function of T, $F_T(n) = P(T <= n)$. To obtain an expression for $F_T(n)$ we define

$$q_m = P(\text{being absorbed at time } m+1 \mid \text{wasn't absorbed by time } m)$$

$$= P(X_{m+1} = K \mid X_m /= K),$$

where /= means unequal.

Hence

$$p_m = P(\text{not being absorbed at time } m+1 \mid \text{wasn't absorbed by time } m)$$

$$= 1 - q_m$$

and

$$P(T > n+1) = 1 - F_T(n+1)$$

$$= \text{PROD}_{m=0}^{n} p_m = \text{PROD}_{m=0}^{n} (1 - q_m),$$

where PROD represents the ordinary arithmetic product.

Thus, $\hat{F}_T(n+1) = 1 - \prod_{m=0}^{n} (1 - \hat{q}_m)$ (2.1)

is an estimator of $F_T(n+1)$, where $\hat{q}_m$ is an estimator of q_m.

We now need to construct $\hat{q}_m$. Define

$$a_j(m) = P(X_m = j)$$

$$r_{jK}(m) = P(X_{m+1} = K \mid X_m = j).$$

We may then write

$$q_m = P(X_{m+1} = K \mid X_m \neq K)$$

$$= \sum_{j=1}^{K-1} P(X_{m+1} = K, X_m = j) \Big/ \sum_{h=1}^{K-1} P(X_m = h)$$

$$= \sum_{j=1}^{K-1} P(X_{m+1} = K \mid X_m = j) P(X_m = j) \Big/ \sum_{h=1}^{K-1} P(X_m = h)$$

$$= \sum_{j=1}^{K-1} a_j(m) r_{jK}(m) \Big/ \sum_{h=1}^{K-1} a_h(m) \qquad (2.2)$$

Recalling that the data available to estimate q_m are statistically independent sample paths, we compute the following quantities:

$N_j(m)$ = number of individuals in state j at time m

$N_{jh}(m)$ = number of individuals in state j at time m and in state h at time $m+1$.

Now, if N is the total number of sample paths then, for each m, we have $\sum_{j=1}^{K} N_j(m) = N$; i.e., we assume that there are no withdrawals

from the study. At this point, we develop the LT and MC estimators of the TTA by computing estimators of q_m based on LT and MC methodology.

2.1 The Life Table Estimator

This estimator results from making no assumptions about the probabilistic structure of X. In particular, we define the estimators

$$\hat{a}_j(m) = N_j(m) / N$$

$$\hat{r}_{jh}(m) = N_{jh}(m) / N_j(m).$$

From these we use expression (2.2) to obtain

$$q_m = \frac{\sum_{j=1}^{K-1} \hat{a}_j(m)\, \hat{r}_{jK}(m)}{\sum_{h=1}^{K-1} \hat{a}_h(m)}$$

$$= \frac{\sum_{j=1}^{K-1} (N_j(m) / N)(N_{jK}(m) / N_j(m))}{\sum_{h=1}^{K-1} N_h(m) / N}$$

$$= \frac{\sum_{j=1}^{K-1} N_{jK}(m)}{\sum_{h=1}^{K-1} N_h(m)} \qquad (2.3)$$

In life table terminology, Chiang (1968), $d_m = \sum_{j=1}^{K-1} N_{jK}(m)$ is the number "dying" (i.e., being absorbed, contracting the condition)

in the interval $[m, m+1)$ and $N_m = \sum_{j=1}^{K-1} N_j(m)$ is the number "alive" at time m. Thus, the estimator of F_T obtained by substituting the estimators of q_m given in (2.3) into expression (2.1) is the complement of the conventional LT survivorship estimator; namely

$$\hat{F}_T(n+1) = 1 - \prod_{m=0}^{n} (1 - d_m / N_m). \qquad (2.4)$$

2.2 The Markov Chain Estimator

Whereas we imposed no assumption on the probabilistic structure of X when deriving the LT estimator, we now assume that X is a stationary Markov chain. That is, we assume that the movement of the process among the states is governed by the transition probability matrix $P = ((p_{ij}))$ where

$$p_{ij} = P(X_n = j \mid X_{n-1} = i).$$

Note that since the underlying process X is stationary these probabilities do not depend on n.

In order to obtain the MC TTA distribution, we need consider the following MC probabilities: (i) the n-step transition probabilities given by

$$p_{ij}(n) = P(X_{m+n} = j \mid X_m = i)$$

and (ii) the first passage time probabilities given by

$$f_{ij}(n) = P(X_n = j, X_m \neq j, 1 \leq m \leq n-1 \mid X_0 = i).$$

These probabilities are related through the well-known expression

$$p_{ij}(n) = \sum_{m=1}^{n} f_{ij}(m)\, p_{jj}(n-m),$$

Thus, we obtain first passage time probabilities iteratively from n-step transition probabilities; we derive the latter from the transition probabilities by making use of the fact that $p_{ij}(n)$ is the (i,j)-th element of the n-th power of P.

Using these quantities Shachtman et al. (1978) show that the MC TTA distribution is given by

$$F_T(n+1) = \sum_{j=1}^{K-1} a_j(0)\, p_{jK}(n+1). \qquad (2.5)$$

We substitute estimates of $p_{jK}(n+1)$ into the above equations to obtain the MC TTA estimates. The $p_{jK}(n+1)$ estimates in turn derive from the maximum likelihood estimators of the p_{ij}'s by employing the previously presented relationship. Finally, the maximum likelihood estimator of p_{ij} is given by

$$\hat{p}_{ij} = N_{ij}/N_i .$$

where

$$N_{ij} = \sum_m N_{ij}(m)$$

$$N_{i.} = \sum_h N_{ih}$$

and

$N_{ij}(m)$ is as previously defined.

3. RELATING THE LIFE TABLE AND MARKOV CHAIN ESTIMATORS

3.1 Analytic Comparison

In this section we provide mathematical verification that when the underlying process X is a stationary MC, the LT and MC estimators

are equivalent in the sense that both estimate the TTA distribution developed at the beginning of the last section. This is done by showing that the functional form of the LT estimator leads to the MC estimator when the Markov assumption is imposed.

THEOREM: Assume that the stochastic process governing movement among a set of states containing an absorbing state is a stationary Markov chain. Then the life table time to absorption distribution, derived without benefit of the Markov assumptions, is equivalent to the time to absorption distribution computed directly from the Markov chain.

PROOF: From expressions (2.1) and (2.2) the LT TTA distribution is given by

$$F_T(n+1) = 1 - \prod_{m=0}^{n} (1-q_m)$$

where

$$q_m = \sum_{j=1}^{K-1} a_j(m)\, r_{jK}(m) \Big/ \sum_{h=1}^{K-1} a_h(m).$$

Stationarity implies that $r_{jK}(m)$ does not depend on m and, furthermore, is equivalent to the MC transition probability p_{jK}; thus q_m may be written as

$$q_m = \sum_{j=1}^{K-1} a_j(m)\, p_{jK} \Big/ \sum_{h=1}^{K-1} a_h(m). \qquad (2.6)$$

Consider the n-step transition probability $p_{ij}(n)$ and, by convention, define

$$p_{ij}(0) = 1 \quad i = j$$
$$= 0 \quad i \neq j.$$

If we further assume, without loss of generality, that there is no chance of being absorbed initially (i.e., $a_K(0) = 0$), then

$$a_j(m) = \sum_{i=1}^{K-1} a_i(0)\, p_{ij}(m), \text{ and equation (2.6) becomes}$$

$$q_m = \sum_{j=1}^{K-1} \sum_{i=1}^{K-1} a_i(0)\, p_{ij}(m)\, p_{jK} \Big/ \sum_{h=1}^{K-1} \sum_{k=1}^{K-1} a_k(0)\, p_{kh}(m).$$

Denote the denominator by b_m $(m = 1, 2, \ldots)$. Then

$$p_m = 1 - q_m = \sum_{j=1}^{K-1} \sum_{i=1}^{K-1} a_i(0)\, p_{ij}(m)\, (1 - p_{jK}) / b_m .$$

Using $1 - p_{jK} = \sum_{h=1}^{K-1} p_{jh}$ and interchanging the summations yields

$$p_m = \sum_{h=1}^{K-1} \sum_{i=1}^{K} a_i(0) \sum_{j=1}^{K-1} p_{ij}(m)\, p_{jh} / b_m .$$

But,

$$\sum_{j=1}^{K} p_{ij}(m)\, p_{jh} = p_{ih}(m+1)$$

and, since $p_{Kh} = 0$ for $h \neq K$, it follows that

$$p_m = \sum_{h=1}^{K-1} \sum_{i=1}^{K-1} a_i(O)\, p_{ih}^{(m+1)} / b_m$$

$$= b_{m+1} / b_m .$$

Thus, using $p_m = 1 - q_m$, F_T is expressible as

$$F_T(n+1) = 1 - \prod_{m=0}^{n} (1-q_m)$$

$$= 1 - (b_{n+1} / b_0)$$

$$= 1 - \frac{\sum_{i=1}^{K-1} \sum_{h=1}^{K-1} a_i(O)\, p_{ih}^{(n+1)}}{\sum_{j=1}^{K-1} a_j(O)} .$$

Since we have assumed $a_K(O) = 0$, it follows that $\sum_{j=1}^{K-1} a_j(O) = 1$ and

$$F_T(n+1) = 1 - \sum_{i=1}^{K-1} \sum_{h=1}^{K-1} a_i(O)\, p_{ih}^{(n+1)}$$

$$= 1 - \sum_{i=1}^{K-1} a_i(O)\,(1 - p_{iK}^{(n+1)})$$

$$= \sum_{i=1}^{K-1} a_i(O)\, p_{iK}^{(n+1)} .$$

This, however, is the MC TTA distribution as given in equation (2.5).

###

3.2 Numerical Comparison

Having verified that the theoretical construct being investigated is the same for both estimation procedures, we now consider the situation where a researcher has a set of sample paths available for analysis and must determine which methodology to employ. If

one has reservations about depicting the underlying process as a stationary MC then there is no choice to make; the LT estimator is robust to this assumption since it does not use it. /1/ If, however, one feels confident about the MC assumption the decision is no longer deterministic. Intuitively, since it correctly incorporates the structure of the process, we feel that the MC estimator should be superior.

This intuition is not easily quantifiable. Because of analytic difficulties in obtaining standard errors of n-step transition probabilities, there do not exist closed form expressions to construct confidence intervals for the MC estimator; hence we are unable to use this traditional criterion to compare the estimators. We have, however, performed a simulation study whose results support our intuition.

In this study we hypothesized several transition matrices and, with each, associated several initial distributions for a stationary MC with five states, one of which was absorbing. The various transition matrices were such that some contained relatively uniform transition probabilities and some contained widely varying transition probabilities. Likewise the selection of initial distributions encompassed vastly differing situations - from uniform over the four transient states to the situation where all mass was concentrated on one of those transient states. Retaining the notation of the previous section, let $F(n) = P(T <= n)$ where T is the time to absorption random variable being studied.

For each combination of transition matrix and initial distribution we employed the following algorithm:

1) Simulate 100 sample paths, each path of length 50 time units.

2) Using these sample paths

a. compute $(\hat{F}_{LT}(n) : n = 1,2,\ldots,50)$, the LT estimator of F.

b. estimate the transition matrix and, using that estimated matrix, compute $(\hat{F}_{MC}(n) : n = 1,2,\ldots,50)$, the MC estimator of F.

3) Repeat steps 1 and 2 one hundred times.

4) Compute (F(n) : n = 1,2,...,50), the true distribution.

Thus we had 100 independent LT estimates of F and 100 independent MC estimates of F, based on 10,000 sample paths. Next we computed, for each n,

1) the mean LT estimate, $\bar{F}_{LT}(n) = (1/100)\ \text{SUM}\ \hat{F}_{LT}(n)$

2) the sample variance of the 100 LT estimates, $S^2_{LT}(n)$

3) the first and ninth deciles of the 100 LT estimates.

4) the mean MC estimate, $\bar{F}_{MC}(n) = (1/100)\ \text{SUM}\ \hat{F}_{MC}(n)$

5) the sample variance of the 100 MC estimates, $S^2_{MC}(n)$

6) the first and ninth deciles of the 100 MC estimates.

Comparing these statistics we discovered little difference between $\bar{F}_{LT}(n)$ and $\bar{F}_{MC}(n)$; see Table 1 for the results from one transition matrix and one initial distribution. Results based on other matrices and other initial distributions did not vary qualitatively from the one in the table. Furthermore, both were close to F(n). This suggests little

difference between the estimates. However, it was usually, although not always, the case that $S^2_{MC}(n) < S^2_{LT}(n)$ and that the inner decile range of the MC estimates was less than the inner decile range of the LT estimates; see Figure A. Since these conclusions held for all transition matrices and for all initial distributions, these observed results do not appear to be an artifact of the particular matrix and/or initial distribution simulated. Thus for each n, a given $\hat{F}_{MC}(n)$ is apparently closer to the true F(n) than is the corresponding $\hat{F}_{LT}(n)$.

Table 1, Figure A

---

4. DISCUSSION

Based on the above results there appears to be little practical numerical difference in results between these analytic approaches. Although initially surprising this is understandable as both procedures are maximum likelihood, although with respect to different underlying assumptions. There are, however, several advantages associated with the MC approach.

The first is that by modeling the entire process, rather than just the initial and outcome state, we obtain significantly greater insight which is useful when interpreting the results. For example, assume that we compare the TTA distributions for two cohorts and find them to differ. If we used MC techniques further investigation of the transition matrix and/or n-step transition probabilities may help to explain the difference. If, however, we used LT techniques,

no such additional help is available. This additional strength of the MC approach is available because we model more states of the process when obtaining the TTA distribution. When employing LT techniques, on the other hand, we omit interim states as consideration of only the outcome state is sufficient. (Note that this means we do not need complete path information to construct a life table and thus the LT estimator exerts a lesser data demand than does the MC estimator.)

Another advantage is that the MC methodology easily lends itself to sensitivity analysis whereas the LT methodology does not. Having estimated the transition matrix, we may postulate changes in (a subset of) the transition probabilities and compute an altered TTA distribution. A comparison of this distribution with the original TTA distribution will assess the extent of the original distribution's dependence on the altered probabilities. There is no apparent analogue to this in LT investigations. Also, any measurements on the MC which may be derived as functions of the transition probabilities (known as parameterizations) may be subjected to straightforward sensitivity analyses.

In addition since the MC allows incorporation of intervening variables, researchers may evaluate potential intervention strategies. Using the methodology of Shachtman, Schoenfelder and Hogue (1980), one may employ the MC to assess what effect would result on TTA if one alters visit patterns to (a subset of the) states, for example, the effect of an intervention which explicitly prevents the process from entering one or more of its states. Shachtman et al. (1976), 1980) provide an example of an intervention where intervening variables are contraceptive states in biological modeling of women's reproductive processes. Again, such an analysis is not feasible using a life table.

Still another area of superiority of the MC involves prediction. Let N be the length of the longest observed interval before absorption. Using the LT estimator it follows that $\hat{F}_{LT}(n) = \hat{F}_{LT}(N)$ for all $n >= N$; that is, we cannot effectively extrapolate into the future. Such a restriction does not exist for the MC estimator as it is not as severely limited by the lengths of the observed sample paths; thus we may use the MC to make predictions into the future. We must, however, exercise care in interpreting such predictions as this situation is similar to that of using a fitted regression line in a domain where there are no observations on the independent variable. As long as we feel comfortable with the prerequisite assumptions, especially stationarity, this prediction into the future provides a reasonable estimate of the desired distribution.

A final, and possibly most important, benefit associated with MC analysis is that it offers the researcher more flexibility than does LT methodology. Consider the (not unusual) situation where one has one data set and desires to address several research questions, each requiring the determination of a TTA type distribution function. After the investigator has estimated the MC transition matrix, all of the TTA distributions are readily obtainable by computing functions of the transition probabilities. Per contra, if the investigator employed LT methodology he would need to construct a separate life table for each research questions.

In summary, we have considered two statistical techniques, one based on LT technology and the other on MC techniques, for estimating the cumulative probability distribution function of the time required for a stochastic process to reach an absorbing state. We have shown that both analytic techniques estimate the same theoretical distribution;

furthermore, results of a simulation study indicate that the techniques perform equally well and that there is apparently little practical numerical difference in results between them. Yet based on the qualitative considerations just discussed, we feel that if the underlying assumptions are justified the MC methodology should be employed. This is not just because it will produce an estimate which is apparently closer to "truth" than would be the LT estimate, but rather because it (the MC methodology) offers the researcher greater flexibility and the opportunity of gaining greater insight into the data.

FOOTNOTES

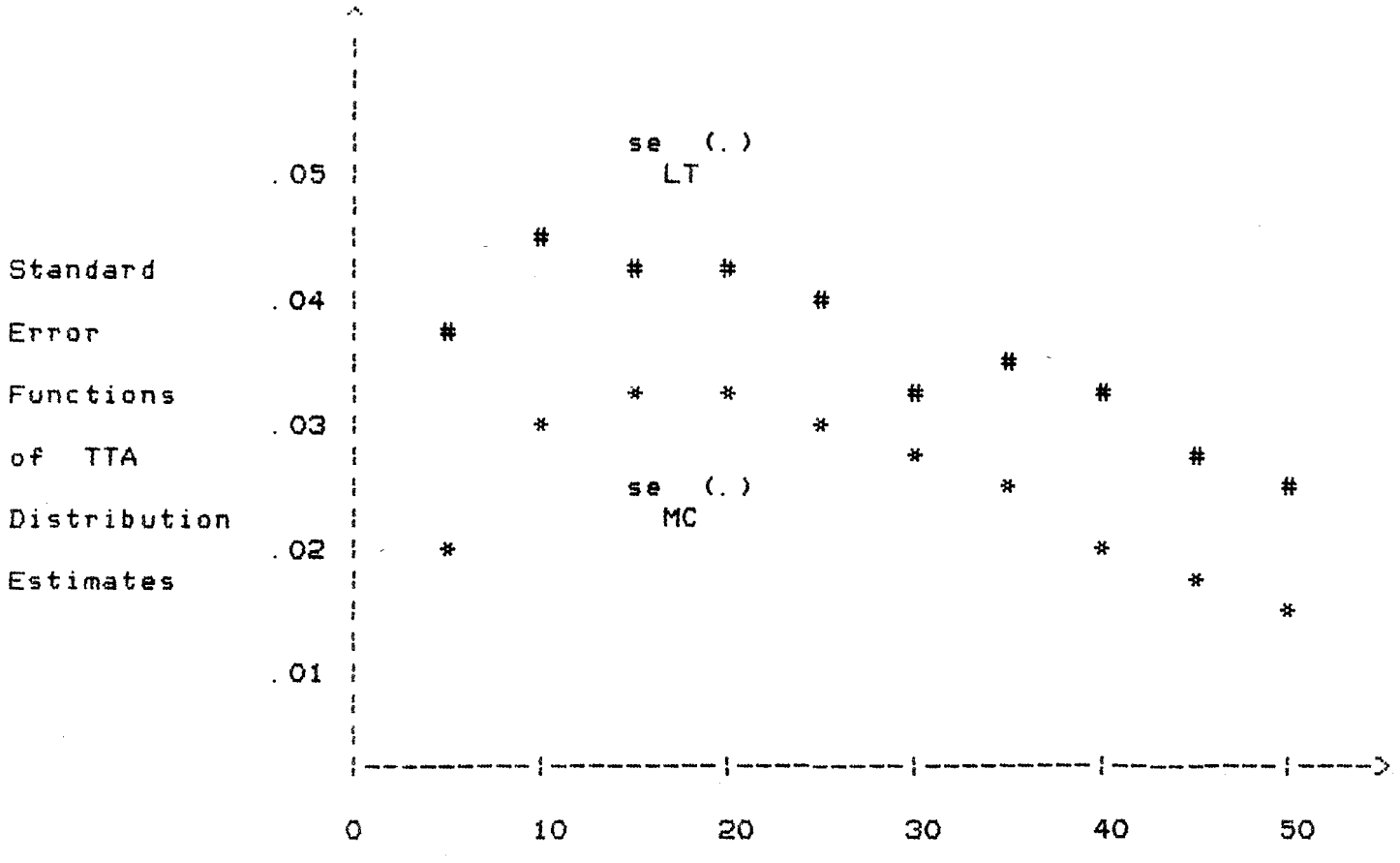
/1/ Frequently a redefinition of the basic state space for the chain will yield a model which is acceptably close to satisfying the Markov and stationarity properties.

TABLE 1. RESULTS OF SIMULATION STUDY OF TIME TO ABSORPTION DISTRIBUTION *

MONTH n	MEAN MC ESTIMATOR $\bar{F}_{MC}(n)$	TRUE TTA DISTRIBUTION $F(n)$	MEAN LT ESTIMATOR $\bar{F}_{LT}(n)$
5	0.246	0.246	0.242
10	0.433	0.434	0.430
15	0.573	0.575	0.573
20	0.678	0.681	0.676
25	0.758	0.760	0.758
30	0.817	0.820	0.816
35	0.862	0.865	0.862
40	0.896	0.899	0.897
45	0.921	0.924	0.923
50	0.940	0.943	0.942

* based on 100 independent MC and LT estimates, each of which is based on 100 independent sample paths

FIGURE A. COMPARISON OF STANDARD ERROR FUNCTIONS FOR TTA DISTRIBUTION MC AND LT ESTIMATORS



$se_{MC}(n)$ = standard error of $\bar{F}_{MC}(n)$

$se_{LT}(n)$ = standard error of $\bar{F}_{LT}(n)$

REFERENCES

Chiang, Chin Long, (1968). INTRODUCTION TO STOCHASTIC PROCESSES IN BIOSTATISTICS. New York, John Wiley and Sons, Inc.

Shachtman, R.H. and Hogue, C.J. (1976). "A Markov Chain Model for Events Subsequent to Induced Abortion." OPERATIONS RESEARCH 24: 916-932.

Shachtman, R.H., Schoenfelder, J.R. and Hogue, C.J. (1978). "Using a Stochastic Model to Investigate Time to Absorption Distributions." (To appear in OPERATIONS RESEARCH)

Shachtman, R.H., Schoenfelder, J.R. and Hogue, C.J. (1980). "Conditional Rate Derivation in the Presence of Intervening Variables Using a Stochastic Model." Institute of Statistics Mimeo Series No. 1292 Department of Biostatistics, University of North Carolina, Chapel Hill. (Submitted to OPERATIONS RESEARCH)