NONPARAMETRIC ESTIMATION OF STAGE TRANSITION TIME
FROM STAGE FREQUENCY DATA

by

Jeffrey S. Pontius

B.A., Millersville University, 1976
M.S., North Dakota State University, 1982

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1987

Approved by:

John E. Boyer, Jr.
Major Professor

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

With great pleasure I would like to here recognize individuals who have contributed significantly to my statistical experience. Foremost, I lovingly thank my wife, Rhoda, for her understanding and support of my academic pursuits and my "course correction" into Statistics. I sincerely thank Dr. John Boyer for his patience, guidence, interest and interaction in the exploration and preparation of this Report and for his support throughout my stay in the Department. I am indebted to Dr. George Milliken for his support and counsel in my career goals and for his consulting and statistical philosophy. To Dr. Stephen Welch I extend my thanks for expanding my narrow horizons into the universe of biological modeling. Also, I would like to recognize Dr. Paul Nelson for his enlightening discussions on statistics, especially double-censoring of the $p_{i,s}$. Retha Parker provided me valuable assistance with WORDMARC. Last (and, of course, not least) I owe Cassie my thanks for her attentive, positive moods.

1. INTRODUCTION

Consider an organism that displays observable stages (0,1,...,A) throughout its lifetime. Two examples are: (1) a holometabolous insect displays egg, larvae, pupa and adult stages (see Ross, 1965) and (2) an annual herbaceous angiosperm displays seed, seedling, vegetative, flowering and senescent stages (see Wilson, Loomis and Steeves, 1971). Let $(t_i)_{i=0}^{F}$ be an increasing sequence of fixed sample points in time (= sample times) such that a cohort of organisms (i) begins in stage 0 at t_0 and (ii) ends in stage A at t_F. At each t_i a sample of n_i (n_i = 1, 2, ...) organisms is selected and the stage of each organism is determined and recorded. Assume that failures (deaths) do not occur and that the samples are selected from an infinitely large population.

The above experiment is applicable in two cases. In Case I, each organism is observed at each t_i. Hence, n_i= N where N is the total number of organisms in the cohort. In Case II, the stage of the organism can only be determined by sacrifice; i.e.: determined by destruction of the organism or the habitat that is necessary for the organism's survival. For example, (1) an adult female insect is sacrificed to determine its ovarian stage and (2) a host (habitat) is destroyed to determine the developmental stage of a parasite (see Ross, 1965). Thus an independent subset n_i of the cohort is observed at each t_i.

Data resulting from such an experiment are tabulated as in Table 1, where $n_{i,s}$ is the number of organisms observed in stage s at t_i. Note that $n_i = \sum_{s=0}^{A} n_{i,s}$. Conditions (i) and (ii) on $(t_i)_{i=0}^{F}$ imply $n_0 = n_{0,0}$ ($n_{0,s} = 0$, $0 < s \le A$) and $n_F = n_{F,A}$ ($n_{F,s} = 0$, $0 \le s < A$). Hence, the estimators to be discussed are appropriate only when all organisms in the cohort begin in stage 0 and end in stage A.

Table 1. Format of stage frequency data.

sample	stage						total
time	0	1	...	s	...	A	no.
$t_0 = 0$	$n_{0,0}$	0	...	0	...	0	n_0
t_1	$n_{1,0}$	$n_{1,1}$	...	$n_{1,s}$	...	$n_{1,A}$	n_1
t_2	$n_{2,0}$	$n_{2,1}$	...	$n_{2,s}$	...	$n_{2,A}$	n_2
⋮	⋮	⋮		⋮		⋮	⋮
t_i	$n_{i,0}$	$n_{i,1}$	...	$n_{i,s}$	...	$n_{i,A}$	n_i
⋮	⋮	⋮		⋮		⋮	⋮
t_F	0	0	...	0	...	$n_{F,A}$	n_F

Given data from such an experiment, a clear interest is to provide estimates of the parameters (mean, variance,...) of the distribution of the time to a particular stage s given the time in any other stage $s' < s$ ($0 \leq s' < A$). Previous research on stage frequency data centered on estimation of survival or mortality rates (Bellows, Ortiz, Owens and Huddleston, 1982; Birley, 1977; Kiritani and Nakasuji, 1967; Manly, 1974). Estimation of stage recruitment was researched by Kobayashi (1968). Estimators of mean stage duration have been proposed by Boyer and Deaton (1984), Manly (1976, 1977) and Mills (1981). Mills (1981) estimated mean stage duration time based on the arithmetic means of recruitment and stage frequencies, scaled by a "shift of mean" factor to account for stage mortality. Manly (1976, 1977) estimated mean stage duration time by applying a trapezoid approximation to the observed stage "frequency estimates." Variance of mean stage duration time was estimated using linear regression sequentially on three frequency estimates (analogous to a moving average). Manly (1985) extended his methodology to data with left or right censoring of sample times. Boyer and Deaton (1984) estimated mean time to stage s in Case II by first constructing a survival cumulative distribution function (cdf) based on the proportion of organisms not yet attaining stage s and then applying Riemann sums to estimate the mean time to stage s. This report is an extension of Boyer and Deaton (1984), so we review their approach next.

Let the random variable $T_s \in [0, \infty[$ be the time to stage s for an organism. Let T_s have the cdf F_s. The survival function for stage s is $P(T_s > t) = G_s(t) = 1 - F_s(t)$. For each t_i, $0 \leq i \leq F$, let $p_{i,s} = P(T_s > t_i) = G_s(t_i)$. $G_s(t_i)$ is the probability that an organism reaches stage s after time t_i and, equivalently, $G_s(t_i)$ is the probability that an organism will not be in stage s by t_i. Define an estimator of $p_{i,s}$ by

$$\hat{p}_{i,s} = \frac{1}{n_i} \sum_{j=0}^{s-1} n_{i,j}$$

(ie: the proportion of organisms in the sample not yet in stage s by time t_i). The quantity $\sum_{j=0}^{s-1} n_{i,j}$ has a binomial($p_{i,s}$, n_i) distribution and $\hat{p}_{i,s}$ is the "usual" unbiased estimator of $p_{i,s}$ (see, for example, Mood, Graybill and Boes, 1974). The $\hat{p}_{i,s}$ would be uniformly minimum variance unbiased if $\hat{p}_{i,s}$ contained information censored from above and below t_i.

Since $E(T_s) = \int_0^\infty G_s(t)dt$ and $G_s(t)$ is monotonic nonincreasing, we can estimate $E(T_s)$ by the Riemann sums $U = \sum_{i=0}^{F-1} G(t_i)(t_{i+1} - t_i)$ and $L = \sum_{i=0}^{F-1} G(t_{i+1})(t_{i+1} - t_i)$, where U (L) is an upper (lower) bound and $(t_i)_{i=0}^{F}$ is the partition. Averaging U and L results in an approximate expression for the mean time to stage s

$$E(T_s) \approx \frac{1}{2} \sum_{i=0}^{F-1} \left[G_s(t_i) + G_s(t_{i+1})\right](t_{i+1} - t_i). \tag{1.2}$$

Substituting $\hat{p}_{i,s}$ for $G_s(t_i)$ and $\hat{p}_{i+1,s}$ for $G_s(t_{i+1})$ and noting that $t_0 = 0$, $\hat{p}_{0,s} = 1$ and $\hat{p}_{F,s} = 0$, an estimator of $E(T_s)$ is

$$\hat{E}(T_s) = \frac{1}{2}t_1 + \frac{1}{2}\sum_{i=1}^{F-1} \hat{p}_{i,s}(t_{i+1} - t_{i-1}). \tag{1.3}$$

Thus $\hat{E}(T_s)$ is approximated as though $G_s(t)$ is a trapezoid in each $[t_i, t_{i+1}]$ (see Fig. 1). To estimate the mean time to stage s from some stage s' take the differences of $\hat{p}_{i,s}$ and $\hat{p}_{i,s'}$ as

$$\hat{E}(T_s - T_{s'}) = \frac{1}{2}\sum_{i=1}^{F-1} (\hat{p}_{i,s} - \hat{p}_{i,s'})(t_{i+1} - t_{i-1}). \tag{1.4}$$

If $s' = s - 1$ then $\hat{E}(T_s - T_{s'})$ is the mean duration time for stage s'. The $\hat{p}_{i,s}$ are independent for different t_i (assuming that the population size is infinitely large), so an estimator of the variance of $\hat{E}(T_s)$ is

$$\hat{\text{Var}}(\hat{E}(T_s)) = \frac{1}{4}\sum_{i=1}^{F-1} \frac{1}{n_i}\left[\hat{p}_{i,s}(1 - \hat{p}_{i,s})\right](t_{i+1} - t_{i-1})^2 \tag{1.5}$$

because $\text{Var}(\hat{p}_{i,s}) = \frac{1}{n_i}[p_{i,s}(1 - p_{i,s})]$. Similarly, an estimator of the variance of $\hat{E}(T_s - T_{s'})$ is

$$\hat{\text{Var}}(\hat{E}(T_s - T_{s'})) = \frac{1}{4}\sum_{i=1}^{F-1} \frac{1}{n_i}\left[(\hat{p}_{i,s} - \hat{p}_{i,s'})(1 - (\hat{p}_{i,s} - \hat{p}_{i,s'}))\right] \times (t_{i+1} - t_{i-1})^2. \tag{1.6}$$

Boyer and Deaton also proved that the absolute bias of $\hat{E}(T_s) \le \frac{1}{2}\max(t_{i+1} - t_i)$ and so $MSE(\hat{E}(T_s)) \le \frac{1}{4}(\max(t_{i+1} - t_i))^2 + Var(\hat{E}(T_s))$. Note that $Var(\hat{E}(T_s))$ is decreased by increasing n_i (ie: taking more samples at each t_i), and bias is decreased by increasing the number of t_i in $[t_0, t_F]$ (ie: refining the partition $(t_i)_{i=0}^{F}$, which is equivalent to sampling more frequently).

In this report, we extend the approach of Boyer and Deaton. In Section 2, we propose two estimators for $Var(T_s)$ and prove two relational properties about the estimators. Section 3 contains an entomological example to demonstrate the calculations of the estimates. A computer program to calculate the estimates is in Section 4. A comparison of the estimators with parameter values from five survival distributions using simulation is presented in Section 5.

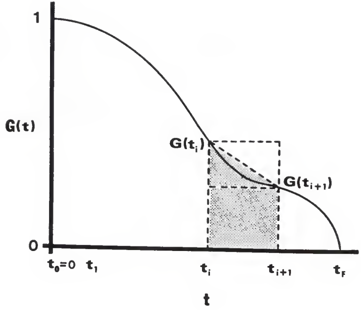

Figure 1. Geometric representation of the calculation of $E(T_s)$.

2. ESTIMATORS OF VAR(T_s)

We propose two estimators of Var(T_s), $\mathrm{Var}_T(T_s)$ and $\mathrm{Var}_L(T_s)$, and prove two relational properties. Var(T_s) may be defined as

$$\mathrm{Var}(T_s) = 2\int_0^\infty t(1 - F_s(t))dt - [E(T_s)]^2 \qquad (2.1)$$

(see Mood, Graybill and Boes, 1974). The crucial estimation is of $E(T_s^2) = 2\int_0^{t_F} tG_s(t)dt$.

2.1 $\mathrm{Var}_T(T_s)$: Trapezoid Method

Let

$$E(T_s^2)_{\Delta t} = 2\int_{t_i}^{t_{i+1}} tG_s(t)dt \qquad (2.2)$$

(ie: the contribution to the second moment of T_s on some $\Delta t = [t_i, t_{i+1}]$). Since $G_s(t)$ is monotonic nonincreasing on Δt,

$2G_s(t_i)\int_{t_i}^{t_{i+1}} tdt \geq 2\int_{t_i}^{t_{i+1}} tG_s(t)dt \geq 2G_s(t_{i+1})\int_{t_i}^{t_{i+1}} tdt$, where $G_s(t_i) = \sup\{G_s(t) \mid t_i \leq t \leq t_{i+1}\}$ and $G_s(t_{i+1}) = \inf\{G_s(t) \mid t_i \leq t \leq t_{i+1}\}$. So $G_s(t_i)(t_{i+1}^2 - t_i^2) \geq 2\int_{t_i}^{t_{i+1}} tG_s(t)dt \geq G_s(t_{i+1})(t_{i+1}^2 - t_i^2)$.

Taking the average of left and right sides we obtain

$E_T(T_s^2)_{\Delta t} \approx \frac{1}{2}(G_s(t_i) + G_s(t_{i+1}))(t_{i+1}^2 - t_i^2)$. So

$$E_T(T_s^2) \approx \frac{1}{2}\sum_{i=0}^{F-1}(G_s(t_i) + G_s(t_{i+1}))(t_{i+1}^2 - t_i^2) \qquad (2.3)$$

and

$$\mathrm{Var}_T(T_s) = E_T(T_s^2) - [E(T_s)]^2 \qquad (2.4)$$

where $E(T_s)$ is (1.2). Thus an estimator of $\mathrm{Var}(T_s)$ is

$$\hat{\mathrm{Var}}_T(T_s) = \frac{1}{2}\sum_{i=0}^{F-1}(\hat{p}_{i,s} + \hat{p}_{i+1,s})(t_{i+1}^2 - t_i^2) - [\hat{E}(T_s)]^2 \qquad (2.5)$$

where $\hat{E}(T_s)$ is (1.3). Expanding (2.5) did not result in a simpler form, so we recommend (2.5) in calculating $\hat{\mathrm{Var}}_T(T_s)$.

2.2 $\mathrm{Var}_L(T_s)$: Straight Line Method

Define $\mathrm{Var}(T_s)$ as (2.1) and $E(T_s^2)_{\Delta t}$ as (2.2). Assume $G_s(t)$ is a linear function on Δt. Then

$$G_s(t) = k + ct = G_s(t_i) + \frac{G_s(t_i) - G_s(t_{i+1})}{t_i - t_{i+1}}(t - t_i).$$

$$E_L(T_s^2)_{\Delta t} = 2\int_{t_i}^{t_{i+1}} tG_s(t)dt$$

$$= 2\int_{t_i}^{t_{i+1}} t\left[G_s(t_i) + \frac{G_s(t_i) - G_s(t_{i+1})}{t_i - t_{i+1}}(t - t_i)\right]dt$$

$$= \left[G_s(t_i) - \frac{G_s(t_i) - G_s(t_{i+1})}{t_i - t_{i+1}}t_i\right](t_{i+1}^2 - t_i^2)$$

$$+ \frac{2}{3}\,\frac{G_s(t_i) - G_s(t_{i+1})}{t_i - t_{i+1}}(t_{i+1}^3 - t_i^3), \text{ and}$$

$$E_L(T_s^2) = \sum_{i=0}^{F-1}\left(G_s(t_i) - \frac{G_s(t_i) - G_s(t_{i+1})}{t_i - t_{i+1}}\, t_i\right)(t_{i+1}^2 - t_i^2)$$

$$+ \frac{2}{3}\sum_{i=0}^{F-1}\frac{G_s(t_i) - G_s(t_{i+1})}{t_i - t_{i+1}}(t_{i+1}^3 - t_i^3) \tag{2.6}$$

and

$$\mathrm{Var}_L(T_s) = E_L(T_s^2) - [E(T_s)]^2 \tag{2.7}$$

where $E(T_s)$ is (1.2). Substituting $\hat{p}_{i,s}$ for $G_s(t_i)$ and $\hat{p}_{i+1,s}$ for $G_s(t_{i+1})$ we define an estimator of $\mathrm{Var}(T_s)$ by

$$\hat{\mathrm{Var}}_L(T_s) = \hat{E}_L(T_s^2) - [\hat{E}(T_s)]^2. \tag{2.8}$$

We expand $E_L(T_s^2)$ to obtain a simpler calculation form of $\hat{\mathrm{Var}}_L(T_s)$. Substituting $\hat{p}_{i,s}$ for $G_s(t_i)$ and $\hat{p}_{i+1,s}$ for $G_s(t_{i+1})$, factoring $(t_{i+1}^2 - t_i^2)$ and $(t_{i+1}^3 - t_i^3)$, and canceling appropriate terms in (2.6) we obtain

$$\hat{E}_L(T_s^2) = \sum_{i=0}^{F-1}\Big(\hat{p}_{i,s}(t_{i+1}^2 - t_i^2) + (\hat{p}_{i,s} - \hat{p}_{i+1,s})(t_{i+1} - t_i)t_i$$

$$+ \frac{2}{3}(\hat{p}_{i,s} - \hat{p}_{i+1,s})(t_{i+1,s}^2 + t_{i+1}t_i + t_i^2)\Big)$$

$$= \sum_{i=0}^{F-1}\Big(\frac{1}{3}\hat{p}_{i,s}(t_{i+1}^2 + t_{i+1}t_i - 2t_i^2)$$

$$+ \frac{1}{3}\hat{p}_{i+1,s}(2t_{i+1}^2 - t_{i+1}t_i - t_i^2)\Big)$$

$$\hat{E}_L(T_s^2) = \frac{1}{3}\sum_{i=0}^{F-1}\left(\hat{p}_{i,s}(t_{i+1} + 2t_i) + \hat{p}_{i+1,s}(2t_{i+1} + t_i)\right)(t_{i+1} - t_i). \tag{2.9}$$

We recommend (2.9) in calculating $\hat{\mathrm{Var}}_L(T_s)$.

2.3 Two Relational Properties

Theorem 1. Define $\mathrm{Var}_T(T_s)$ as (2.4) and $\mathrm{Var}_L(T_s)$ as (2.7). Given $(t_i)_{i=0}^F$, $t_i \geq 0$, $t_{i+1} > t_i$ and $G_s(t_i) \geq G_s(t_{i+1})$ then $\mathrm{Var}_T(T_s) > \mathrm{Var}_L(T_s)$.

Proof: Note that proving $\mathrm{Var}_T(T_s) > \mathrm{Var}_L(T_s)$ is equivalent to proving $0 < E_T(T_s^2) - E_L(T_s^2)$. From (2.3) and substituting $G_s(t_i)$ for $\hat{p}_{i,s}$ and $G_s(t_{i+1})$ for $\hat{p}_{i+1,s}$ in (2.9), we obtain $E_T(T_s^2) - E_L(T_s^2)$

$$= \frac{1}{2}\sum_{i=0}^{F-1}[G_s(t_i) + G_s(t_{i+1})](t_{i+1} + t_i)(t_{i+1} - t_i)$$

$$- \frac{1}{3}\sum_{i=0}^{F-1}\left(G_s(t_i)(t_{i+1} + 2t_i) + G_s(t_{i+1})(2t_{i+1} + t_i)\right)(t_{i+1} - t_i)$$

$$= \sum_{i=0}^{F-1}\left(\frac{1}{2}G_s(t_i)(t_{i+1} + t_i) + \frac{1}{2}G_s(t_{i+1})(t_{i+1} + t_i)\right.$$

$$\left.- \frac{1}{3}G_s(t_i)(t_{i+1} + 2t_i) - \frac{1}{3}G_s(t_{i+1})(2t_{i+1} + t_i)\right)(t_{i+1} - t_i)$$

$$= \frac{1}{6}\sum_{i=0}^{F-1}\left(G_s(t_i)(t_{i+1} - t_i) + G_s(t_{i+1})(-t_{i+1} + t_i)\right)(t_{i+1} - t_i)$$

$$= \frac{1}{6}\sum_{i=0}^{F-1}[G_s(t_i) - G_s(t_{i+1})](t_{i+1} - t_i)^2 \geq 0.$$ Thus

$0 < \sum_{i=0}^{F-1} [G_s(t_i) - G_s(t_{i+1})](t_{i+1} - t_i)^2$. Strict inequality holds because $G_s(t)$ is nonincreasing monotonic, $G_s(t_0) = 1$ and $G_s(t_F) = 0$. Thus $Var_T(T_s) > Var_L(T_s)$ completing the proof. □

Theorem 1 holds for the sample estimates $\hat{Var}_T(T_s)$ and $\hat{Var}_L(T_s)$ when the $\hat{p}_{i,s}$ have the same monotonic property. Note that the $\hat{p}_{i,s}$ may not be monotonic because of sampling variation. However, we have always observed that $\hat{Var}_T(T_s) > \hat{Var}_L(T_s)$ in simulations and actual data applications.

In the following Lemma we show that as the number of t_i in $[t_0, t_F]$ increases, $\hat{Var}_T(T_s)$ approaches $\hat{Var}_L(T_s)$.

<u>Lemma 1.</u> Let $(t_i)_{i=0}^{F}$, $t_i \geq 0$, $t_{i+1} > t_i$ and $(n_i)_{i=0}^{F}$ be given. Define $\Delta t = t_{i+1} - t_i$. Then $\hat{Var}_T(T_s) \to \hat{Var}_L(T_s)$ as $\Delta t \to 0$.

Proof: Fix s. From Theorem 1, and substituting $\hat{p}_{i,s}$ for $G_s(t_i)$ and $\hat{p}_{i+1,s}$ for $G_s(t_{i+1})$, we have

$$\hat{Var}_T(T_s) - \hat{Var}_L(T_s) = \hat{E}_T(T_s^2) - \hat{E}_L(T_s^2) = \frac{1}{6}\sum_{i=0}^{F-1}(\hat{p}_{i,s} - \hat{p}_{i+1,s})(\Delta t)^2.$$

Recall that the right side may be less than zero if the $\hat{p}_{i,s}$ are not monotonic nonincreasing. Now,

$$\left|\frac{1}{6}\sum_{i=0}^{F-1}(\hat{p}_{i,s} - \hat{p}_{i+1,s})(\Delta t)^2\right| \leq \left|\frac{1}{6}\max_i(\Delta t)^2\sum_{i=0}^{F-1}(\hat{p}_{i,s} - \hat{p}_{i+1,s})\right|$$

$$= \frac{1}{6}\max_i(\Delta t)^2 \text{ since } \hat{p}_{0,s} = 1 \text{ and } \hat{p}_{F,s} = 0.$$

Thus $\lim\limits_{\Delta t \downarrow 0} \frac{1}{6}\max\limits_i(\Delta t)^2 = 0$ and $\hat{\mathrm{Var}}_T(T_s) \to \hat{\mathrm{Var}}_L(T_s)$ as $\Delta t \to 0$. □

3. AN EXAMPLE

We provide an example of the calculations for $\hat{E}(T_s)$, $\hat{E}(T_s - T_{s'})$, $\hat{Var}(\hat{E}(T_s))$, $\hat{Var}(\hat{E}(T_s - T_{s'}))$, $\hat{Var}_T(T_s)$ and $\hat{Var}_L(T_s)$. The data are from a stage frequency experiment on the insect Lysiphlebus testaceipes (Hymenoptera: Aphidiidae), an endoparasitoid of the aphid, Schizaphis graminum (Homoptera: Aphidae) (Table 2) (unpublished data, J. S. Pontius) (see Hight, Eikenbary, Miller and Starks (1972) for L. testaceipes biology). The $\hat{p}_{i,s}$ for the example calculations are included in Table 2. Since $t_1 = 4$, let $t_0 = 0$ and $\hat{p}_{0,s} = 1$.

Table 2. Stage frequency data for L. testaceipes.

sample time (days)	stage 0 egg-larva	1 pupa	2 adult	total no.	$\hat{p}_{i,1}$	$\hat{p}_{i,2}$
t_0=0	-	-	-	-	1	1
4	8	0	0	8	1	1
7	15	0	0	15	1	1
9	3	15	0	18	$.1\underline{6}$	1
11	0	14	0	14	0	1
13	0	2	12	14	0	.1429
15	0	0	2	2	0	0

The estimated mean time to stage 1 (pupa) is $\hat{E}(T_1) = .5(4) + [(1)(7 - 0) + (1)(9 - 4) + .1\underline{6}(11 - 7) + 0 + 0] = 8.\underline{3}$ from (1.3). The estimated standard error of the mean time to stage 1 is $\hat{Var}(\hat{E}(T_1))$

$= .25[\ 0 + 0 + .05\underline{5}(.1\underline{6}(1 - .1\underline{6}))(11 - 7)^2 + 0 + 0 + 0] = .031$ from (1.5). The estimated mean duration time in stage 1 is $\hat{E}(T_1 - T_0) = .5[0 + 0 + 0 + .8\underline{3}(11 - 7) + (1)(13 - 9) + .1429(15 - 11)] = 3.95$ from (1.4). From (1.6), $\hat{Var}(\hat{E}(T_1 - T_0)) = .25[0 + 0 + 0 + .0\underline{5}(.8\underline{3}(1 - .8\underline{3}))(11 - 7)^2 + .071(.1429(1 - .1429))(15 - 11)^2] = .067$. Estimates of moments for variances of time to stage 1 are $\hat{E}_T(T_1^2) = .5[(1 + 1)(4^2 - 0) + (1 + 1)(7^2 - 4^2) + (1 + .1\underline{6})(9^2 - 7^2) + (.1\underline{6} + 0)(11^2 - 9^2) + 0 + 0] = 71$ and $\hat{E}_L(T_1^2) = .\underline{3}\{[(1)(4 + 0) + (1)(8 +)](4 - 0) + [(1)((7 + 8) + (1)(14 + 4)](7 - 4) + [(1)(9 + 14) + .1\underline{6}(18 + 7)](9 - 7) + [.1\underline{6}(11 + 18) + 0](11 - 9) + 0 + 0]\} = 70.\underline{3}$ from (2.3) and (2.9), respectively. Hence, from (2.5) and (2.8), the estimated variances of time to stage 1 are $\hat{Var}_T(T_1) = 1.\underline{5}$ and $\hat{Var}_L(T_1) = .\underline{8}$.

4. A COMPUTER PROGRAM FOR CALCULATING ESTIMATES

We developed a computer program (Appendix A) to calculate $\hat{E}(T_s)$, $\hat{Var}(\hat{E}(T_s))$, $\hat{E}(T_s - T_{s'})$, $\hat{Var}(\hat{E}(T_s - T_{s'}))$ {s' = s - 1}, $\hat{Var}_T(T_s)$ and $\hat{Var}_L(T_s)$ from a set of stage frequency data (see Table 1). The program was coded in the Macro Language of the Statistical Analysis System (SAS) (Allen, 1982), version 82.3. SAS Macro Language requires a minimum of 500 K mamory and the JCL execution card //ƀEXECƀSAS,OPTIONS=MACRO. The source code contains documentation for SAS data set structure, computational algorithms and user required initial values.

The SAS data set structure is documented in the program. The stage frequency table (see Table 1) is structured as a column formatted input data set, excluding the totals column. The program checks that the data set contains (1) nonnegative count data ($n_{i,s} \geq 0$, $0 \leq i \leq F$, $0 \leq s \leq A$) and (2) an increasing sequence of nonnegative sample times $(t_i)_{i=0}^{F}$. Error messages are printed on the SAS log if errors in (1) and/or (2) are detected. If the data do not contain $t_0 = 0$, an algorithm inserts $t_0 = 0$, $n_{0,0} = 1$ and $n_{0,s} = 0$, $0 < s \leq A$, as the first row of the data set.

User-required initial values are in the main program section located at the end of the source code. The initial values are (1) number of stages in the data set (NS), (2) variables corresponding to each stage (COUNTk), (3) a list of character identifiers for the stages, and (4) the unit of measurement for sample times.

The estimates are calculated using matrices in PROC MATRIX. Three sections of output are generated by the program (Table 3). The first section contains the input data set. The variables COUNTk are as initialized in (2) above, k= 1, 2, ..., A+1 (corresponding to $s = 0, 1, \ldots, A$), and the values listed under DAY are sample times. The second section lists the stage identifiers and the estimates $\hat{E}(T_s)$, $\sqrt{[\hat{Var}(\hat{E}(T_s))]}$, $\hat{E}(T_s - T_{s'})$ and $\sqrt{[\hat{Var}(\hat{E}(T_s - T_{s'}))]}$ for $s' = s - 1$. The third section contains the stage identifiers, $\sqrt{[\hat{Var}_T(T_s)]}$ and $\sqrt{[\hat{Var}_L(T_s)]}$. Estimates for the pupal stage in Table 3 correspond to the example calculations in Section 3.

Table 3. Example of computer program output. Estimates correspond to the example in Section 3.

ESTIMATION OF TIME TO AND DURATION OF STACE
FREQUENCY DATA, METHOD OF BOYER AND DEATON.
PROCRAM REVISED: SEPTEMBER, 1986 BY JS PONTIUS.
STUDY: L. TESTACEIPES
UNIT OF TIME MEASUREMENT: DAYS

OBS	DAY	COUNT1	COUNT2	COUNT3
1	4	8	0	0
2	7	15	0	0
3	9	3	15	0
4	11	0	14	0
5	13	0	2	12
6	15	0	0	2

ESTIMATION OF TIME TO AND DURATION OF STACE
FREQUENCY DATA, METHOD OF BOYER AND DEATON.
PROCRAM REVISED: SEPTEMBER, 1986 BY JS PONTIUS.
STUDY: L. TESTACEIPES
UNIT OF TIME MEASUREMENT: DAYS

STACE	TIME TO REACH STACE (E(T(S)))	STD ERROR OF E(T(S))	DURATION TIME E(T(S)) - T(S'))	STD ERROR OF E(T(S) - T(S'))
EGGLARVA	.	.	8.33333	0.175682
PUPA	8.3333	0.175682	3.95238	0.256612
ADULT	12.2857	0.187044	.	.

ESTIMATION OF TIME TO AND DURATION OF STAGE
FREQUENCY DATA, METHOD OF BOYER AND DEATON.
PROGRAM REVISED: SEPTEMBER, 1986 BY JS PONTIUS.
STUDY: L. TESTACEIPES
UNIT OF TIME MEASUREMENT: DAYS

STACE	STD DEVIATION OF T(S) -TRAPEZOID ANALOG-	STD DEVIATION OF T(S) -STRAICHT LINE-
PUPA	1.24722	0.942809
ADULT	1.22057	0.907265

5. SIMULATION UNDER FIVE SURVIVAL DISTRIBUTIONS

We compare the performance of $\hat{E}(T_s)$, $\hat{Var}_T(T_s)$, $\hat{Var}_L(T_s)$ and $\hat{Var}(\hat{E}(T_s))$ to their respective expected values $E(T_s)$ and $Var(T_s)$ (for the first 3 estimators) of 5 survival ($G_s(t)$) distributions under specified sampling conditions for a cohort of organisms with 2 stages, $s = \{0, 1\}$. We selected the uniform, exponential, beta, normal and standard gamma survival distributions to provide a variety of distributional shapes (Table 4).

Table 4. Survival $G_s(t)$, parameter value(s), $E(T_s)$ and $Var(T_s)$ for each survival distribution used in simulations.

survival distribution	parameter value(s)	$E(T_s)$	$Var(T_s)$
uniform: $G_s(t) = (1 - t)I_{[0,1]}(t)$	-	0.5	$1/\sqrt{12}$
exponential: $G_s(t) = e^{-\lambda t}I_{[0,\infty)}(t)$	$\lambda = 1$	1.0	1.0
beta: $G_s(t) = 1 - \int_0^t \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{[0,1]}(x)$	$\alpha = 2.0$ $\beta = 0.5$	0.8	0.046
normal: $G_s(t) = 1 - \int_0^t \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) I_{(0,\infty)}(x)$	$\mu = 3.5$ $\sigma = 1.0$	3.5	1.0
gamma: $G_s(t) = 1 - \int_0^t \frac{1}{\Gamma(\alpha)} x^{\alpha-1}e^{-x}I_{[0,\infty)}(x)$	$\alpha = 2.0$	2.0	2.0

We selected the uniform because we expected $\hat{E}(T_s)$ to closely estimate $E(T_s)$ since the trapezoidal approximation should do well on the uniform's straight line survival distribution (Fig. 2 and see Fig. 1). We expected $\hat{E}(T_s)$ to overestimate $E(T_s)$ from an exponential $G_s(t)$ because the trapezoidal approximation should overestimate the area under the convex exponential survival distribution (Fig. 3). Conversely, we expected $\hat{E}(T_s)$ to underestimate $E(T_s)$ from a concave beta survival distribution (Fig. 4). Because a normal distribution is symmetric about $E(T_s)$, we expected underestimation of the trapezoids at the upper tail and overestimation of trapezoids at the lower tail (Fig. 5). Hence, possibly some cancelling effect would occur. We selected a gamma because it is a widely used survival distribution and exhibits shape asymmetry (Fig. 6).

The variables we used in sampling from each survival distribution were (1) the time interval (Δt) between sample times $(t_i)_{i=0}^{F}$ and (2) the number of organisms (n_i) in stages 0 and 1 'observed' at each t_i. To construct a similar sampling regimen over distributions, Δt was scaled with respect to $Var(T_s)$ of each distribution as

$$\Delta t = c\sqrt{Var(T_s)}, \tag{5.1}$$

where c = 1, 1/2, 1/4 or 1/8. Thus Δt was constant for each simulation (ie: sample times were equidistant). Note that Δt is just a proportion of the standard deviation of each $G_s(t)$. Also, the number of samples (ie: the 'number of organisms observed') per t_i were n_i = 5, 10, 20 or 40 for all t_i, $i = 0,\ldots,F$. For brevity, let $n = n_i$, $i = 0,\ldots,F$. Thus the performance of the estimators was evaluated based on 16 combinations of c and n, for each survival distribution. Fifty 'cohorts' were sampled for each combination of c and n.

The statistics we used to evaluate the performance of the estimators were mean and root mean square of $\hat{E}(T_s)$, $\hat{Var}_T(T_s)$ and $\hat{Var}_L(T_s)$ each and the mean of $\hat{Var}(\hat{E}(T_s))$. Clearly, the sample size for the mean and root mean square of each estimator is 50.

5.1 Simulation Algorithm

We describe the algorithm used in our simulations. We used PROC MATRIX of SAS, version 5, (processed on a NAS 6630) because our computer algorithms to calculate estimates were already coded in SAS (see Section 4). Version 5 was used because the RANBIN binomial random number generator in version 82.3 was defective. We were convinced that RANBIN in version 5 was correct after several tests on RANBIN were performed.

The simulation algorithm is as follows. Algorithm instructions apply to each survival distribution. First intialize $E(T_s)$, $Var(T_s)$, c and n. Next generate the sample times $(t_i)_{i=0}^{F}$. Set $t_0 = 0$. Generate sample times $(t_i)_{i=1}^{F-1}$ by $t_{i+1} = t_i + \Delta t$, $i = 0,\ldots,F-2$, where Δt is determined from (5.1). To ensure that $p_{F,s} = 0$, calculate t_F for uniform and beta survival distributions as $t_F = t_{F-1} + \Delta t$ where $t_{F-1} = \max\{t_i \in [0,1]\}$. Then calculate t_F for exponential, normal and gamma survival distributions as $t_F = t_{F-1} + \Delta t$ where

$t_{F-1} = \max\{t_i \mid P(T_s < t_i) \le 0.9999\}$.

For each t_i, determine a corresponding $p_{i,0}$. Set $p_{0,0} = 1$. The $p_{i,0}$, $i = 1,\ldots,F-1$, for uniform and exponential are determined from survival distributions (see Table 4). The $p_{i,0}$, $i = 1,\ldots,F-1$, for beta, normal and gamma are calculated as $p_{i,0} = 1 - P(T_s < t_i)$ where $P(T_s < t_i)$ are determined from the SAS probability generators PROBBETA, PROBNORM and PROBGAM, respectively. Parameter values for the probability generators are listed in Table 4. Set all $p_{F,0} = 0$.

Now determine the number of organisms 'observed' in stage 0. For each of the 50 cohort simulations do the following. Initialize all $n_{i,0} = 0$ for each survival distribution. Set $n_{0,0} = n$ (ie: all organisms are in stage 0 at $t_0 = 0$). For each t_i, $i = 1,\ldots,F$, randomly select the number of organisms in stage 0 using the binomial random number generator RANBIN with parameters $p_{i,0}$ and n. Stop sampling when the first $n_{i,0} = 0$, $i > 0$. Calculate $n_{i,1} = n - n_{i,0}$, $i = 0,\ldots,F$, to determine the number of organisms that have reached stage 1 by t_i. Calculate $\hat{E}(T_s)$, $\hat{V}ar(\hat{E}(T_s))$, $\hat{V}ar_T(T_s)$ and $\hat{V}ar_L(T_s)$ and store the estimates.

After 50 cohorts have been simulated, calculate the evaluation statistics specified above. Print the evaluation statistics. This ends the algorithm.

5.2 Results and Discussion of Simulations

We present the simulations' results and corresponding discussion in the order $\hat{E}(T_s)$, $\hat{Var}(\hat{E}(T_s))$, $\hat{Var}_T(T_s)$ and $\hat{Var}_L(T_s)$. For each estimator we present its overall performance and then present results pertinent to selected survival distributions. Each sampling regimen will be referenced by c, each sample size per t_i will be referenced by n, and specific combinations of c and n will be referenced by (c ; n). Section 5.4 contains an overall discussion of the simulations.

5.2.1 $\hat{E}(T_s)$

Our overall evaluation of the performance of $\hat{E}(T_s)$ is based on trends in the means and estimated root mean squares ($\hat{RMS}$) of $\hat{E}(T_s)$ (Table 5) and 95% confidence intervals for $E(T_s)$ (Table 6). We constructed 95% confidence intervals for $E(T_s)$ by mean($\hat{E}(T_s)$) ± $1.96\sqrt{[\text{mean}(\hat{Var}(\hat{E}(T_s)))/50]}$ where values of mean($\hat{Var}(\hat{E}(T_s))$) are listed in Table 7 and 50 is the number of simulations for each survival distribution, c and n combination. If $E(T_s)$ is contained in the confidence interval then we consider $\hat{E}(T_s)$ to be a good estimator of $E(T_s)$ under the particular survival distribution, c and n combination. Note that the confidence intervals can also be used to test H_o: $\hat{E}(T_s)$ is an unbiased estimate of $E(T_s)$ vs. H_a: $\hat{E}(T_s)$ is a biased estimate of

$E(T_s)$. If H_o is rejected, the bias can be estimated by $\widehat{\text{bias}}$ = mean($\hat{E}(T_s)$) - $E(T_s)$ from the values in Table 5.

Overall, $E(T_s)$ is contained all confidence intervals for the beta survival distribution (Table 6), $E(T_s)$ is contained in most confidence intervals for the uniform and normal survival distributions, and $E(T_s)$ is contained in less than half of the confidence intervals for the exponential and gamma survival distributions. Thus we conclude that $\hat{E}(T_s)$ is a good estimator of the expected values $E(T_s)$ for the beta and uniform survival distributions, and for (1 ; 5, 10, 20, 40), (1/2 ; 20, 40), (1/4 ; 10, 20, 40) and (1/8 ; 20, 40) for the normal survival distribution. $\hat{E}(T_s)$ appears to be a biased estimator of $E(T_s)$ of the exponential and gamma survival distributions when $c = 1/4$ or $c = 1/8$. Even though the bias of $\hat{E}(T_s)$ should decrease as Δt decreases (see Section 1), this assumes that all t_i have been sampled. Possibly the presence of bias when $c = 1/4$ and $c = 1/8$ for the exponential and gamma survival distributions is because of some t_i (in the right part of the survival distribution) consistently not being sampled in simulations.

For the exponential, normal and gamma survival distributions, the means of $\hat{E}(T_s)$ tended to decrease as c decreased and n increased. However, the means of $\hat{E}(T_s)$ tended to be similar across all (c ; n) for the uniform and beta survival distributions. Means of $\hat{E}(T_s)$ were closer overall to $E(T_s)$ of beta and uniform survival distributions. Conversely, means of $\hat{E}(T_s)$ were farther overall from $E(T_s)$ of the gamma survival distribution. Possibly $\hat{E}(T_s)$ better estimates $E(T_s)$ of survival distributions with shapes similar to the concave beta and uniform survival distributions. But the [0, 1] domains of the uniform and beta survival distributions compared with the [0, ∞[domains of the exponential, normal and gamma survival distributions may have influenced these results. Estimated root mean squares of $\hat{E}(T_s)$ tended to decrease as c decreased and n increased (Table 5).

For the exponential survival distribution, $\hat{E}(T_s)$ overestimated $E(T_s)$ for (1 ; 5, 10, 20, 40) as we expected but $\hat{E}(T_s)$ tended to underestimate $E(T_s)$ for (1/2, 1/4, 1/8 ; 5, 10, 20). For the gamma survival distribution, $\hat{E}(T_s)$ tended to underestimate $E(T_s)$, especially for smaller n and c.

5.2.2 $\hat{Var}(\hat{E}(T_s))$

Boyer and Deaton (1984) concluded that $Var(\hat{E}(T_s))$ would be decreased by taking more samples (ie: increasing n_i) at each t_i. Means of $\hat{Var}(\hat{E}(T_s))$ from the simulations (Table 7) support their conclusion. For each survival distribution and c, means of $\hat{Var}(\hat{E}(T_s))$ decreased as n increased; and for each survival distribution and n, means of $\hat{Var}(\hat{E}(T_s))$ decreased as c decreased. Overall, we conclude that increasing n_i at each t_i and increasing the number of t_i in $[t_0, t_F]$ results in a smaller estimate of $Var(\hat{E}(T_s))$.

5.2.3 $\hat{Var}_T(T_s)$

The $\hat{Var}_T(T_s)$ tended to overestimate $Var(T_s)$ of each survival distribution when c = 1, and for smaller n (usually n = 5 or n = 10) for all other values of c (Table 8). Overall, $\hat{Var}_T(T_s)$ better estimated $Var(T_s)$ for smaller c and larger n under the uniform, beta and normal survival distributions; and for smaller c under the exponential and gamma survival distributions. Means of $\hat{Var}_T(T_s)$ were closer overall to $Var(T_s)$ of the beta survival distribution and farther from $Var(T_s)$ of the exponential and gamma survival distributions. So $\hat{Var}_T(T_s)$ better estimated $Var(T_s)$ of beta and uniform survival distributions. Trends in $\hat{RMS}$ of $\hat{Var}_T(T_s)$ were variable and appear to depend on the particular survival distribution.

For the uniform survival distribution, $\hat{Var}_T(T_s)$ overestimated $Var(T_s)$ when c = 1, (1/4 ; 5) and (1/8 ; 5). For all n when c = 1/2 and for (1/4, 1/8 ; 10, 20, 40) $\hat{Var}_T(T_s)$ reasonably estimated $Var(T_s)$. The $\hat{RMS}$'s of $\hat{Var}_T(T_s)$ tended to decrease as c decreased and n increased.

For the exponential survival distribution, $\hat{Var}_T(T_s)$ varied considerably with changes in n within and across values of c. Means of $\hat{Var}_T(T_s)$ were closer to $Var(T_s)$ for (1 ; 20, 40). For all other (c ; n), $\hat{Var}_T(T_s)$ underestimated $Var(T_s)$. In (1/2 ; 5) and (1/4, 1/8 ; 5, 10) $\hat{Var}_T(T_s)$ underestimated $Var(T_s)$ considerably. The $\hat{RMS}$ of $\hat{Var}_T(T_s)$ decreased as n increased for each c.

For the beta survival distribution, $\hat{Var}_T(T_s)$ reasonably estimated $Var(T_s)$, particularly as c decreased and n increased. The $\hat{RMS}$ of $\hat{Var}_T(T_s)$ decreased as c decreased and n increased.

For the normal survival distribution, $\hat{Var}_T(T_s)$ overestimated $Var(T_s)$ when c = 1. For all other (c ; n), except for (1/4, 1/8 ; 5), $\hat{Var}_T(T_s)$ reasonably estimated $Var(T_s)$. The $\hat{RMS}$ of $\hat{Var}_T(T_s)$ decreased as n increased for each c.

For the gamma survival distribution, $\hat{Var}_T(T_s)$ varied considerably with changes in n within and across values of c. Means of $\hat{Var}_T(T_s)$ were closer to $Var(T_s)$ for (1 ; 5, 10) and (1/2 ; 40). In (1/4 ; 5) and (1/8 ; 5, 10), $\widehat{Var}_T(T_s)$ underestimated $Var(T_s)$ considerably. The $\widehat{RMS}$ of $\widehat{Var}_T(T_s)$ decreased as n increased for each c.

5.2.4 $\widehat{Var}_L(T_s)$

Means of $\widehat{Var}_L(T_s)$ (Table 9) were less than the means of $\widehat{Var}_T(T_s)$ for all survival distributions and (c ; n) (equality is the result of rounding of estimates) as proved in Theorem 1 (see Section 2.3). In general, statements about $\widehat{Var}_T(T_s)$ in Section 5.2.3 pertain to $\widehat{Var}_L(T_s)$. As c decreased and n increased, $\widehat{Var}_L(T_s)$ approached $\widehat{Var}_T(T_s)$ as proved in Lemma 1 (see Section 2.3). The $\widehat{RMS}$'s of both estimators were similar for smaller c (especially c = 1/8). Hence, $\widehat{Var}_L(T_s)$ and $\widehat{Var}_T(T_s)$ give similar estimates of $Var(T_s)$ as the number of t_i in $[t_0, t_F]$ and n_i are increased.

Figure 2. Graph of uniform $G_s(t)$ used in simulations.

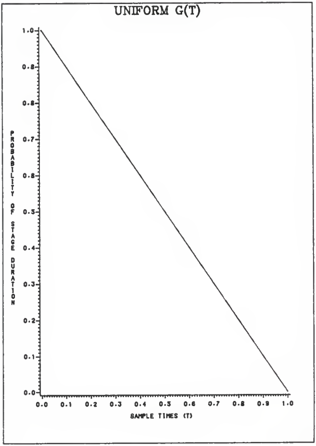

Figure 3. Graph of exponential $G_s(t)$ used in simulations.

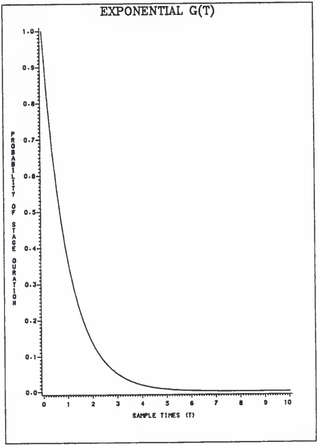

Figure 4. Graph of beta $G_s(t)$ used in simulations.

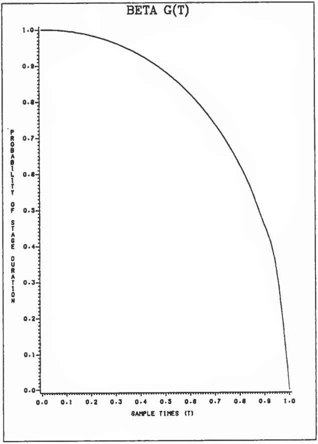

Figure 5. Graph of normal $G_s(t)$ used in simulations.

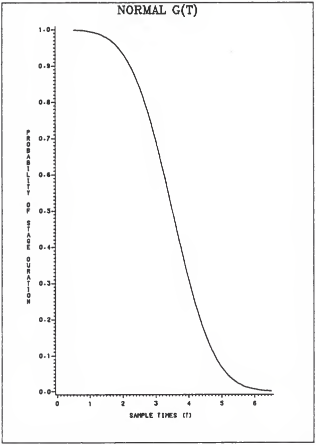

Figure 6. Graph of gamma $G_s(t)$ used in simulations.

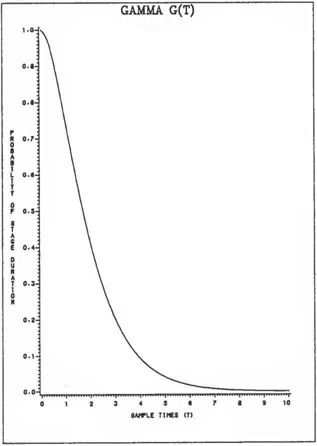

Table 5. $\hat{E}(T_s)$ results from survival distribution simulations. Mean of $\hat{E}(T_s)$ and (root mean square) are listed for each c and n combination under each survival distribution.

DISTRIBUTION:		uniform	exponential	beta	normal	gamma
$E(T_s)$:		.5	1.0	.8	3.5	2.0
c	n					
1	5	.50 (.10)	1.06 (.35)	.80 (.07)	3.45 (.34)	1.88 (.46)
	10	.53 (.09)	1.02 (.21)	.79 (.05)	3.53 (.28)	1.96 (.29)
	20	.51 (.05)	1.06 (.16)	.80 (.04)	3.53 (.17)	2.05 (.24)
	40	.51 (.04)	1.08 (.14)	.80 (.02)	3.53 (.11)	1.98 (.18)
$\frac{1}{2}$	5	.48 (.08)	.83 (.31)	.80 (.05)	3.41 (.28)	1.81 (.39)
	10	.50 (.06)	.95 (.18)	.81 (.04)	3.43 (.21)	1.90 (.25)
	20	.50 (.03)	.99 (.09)	.80 (.03)	3.51 (.12)	1.98 (.19)
	40	.50 (.02)	1.00 (.06)	.80 (.02)	3.50 (.10)	1.99 (.11)
$\frac{1}{4}$	5	.50 (.06)	.81 (.28)	.79 (.03)	3.45 (.22)	1.78 (.33)
	10	.50 (.04)	.90 (.16)	.80 (.02)	3.47 (.10)	1.90 (.19)
	20	.50 (.02)	.92 (.12)	.80 (.02)	3.50 (.07)	1.94 (.13)
	40	.50 (.02)	.99 (.07)	.80 (.01)	3.49 (.07)	1.97 (.10)
$\frac{1}{8}$	5	.44 (.09)	.76 (.29)	.79 (.03)	3.34 (.22)	1.65 (.42)
	10	.49 (.04)	.84 (.20)	.80 (.02)	3.43 (.12)	1.78 (.25)
	20	.50 (.02)	.93 (.10)	.80 (.01)	3.49 (.05)	1.94 (.12)
	40	.50 (.01)	.96 (.05)	.80 (.01)	3.49 (.04)	1.96 (.09)

Table 6. 95% confidence intervals for $E(T_s)$ based on simulations. Confidence intervals are listed for each c and n combination under each survival distribution. * labels confidence intervals that do not contain $E(T_s)$.

DISTRIBUTION:		uniform	exponential	beta	normal	gamma
	$E(T_s)$:	.5	1.0	.8	3.5	2.0
c	n					
1	5	(.48,.53)	(.99,1.13)	(.78,.82)	(3.36,3.54)	(1.77,1.99)*
	10	(.51,.54)*	(.97,1.07)	(.78,.80)	(3.47,3.59)	(1.88,2.04)
	20	(.50,.52)	(1.02,1.10)*	(.79,.81)	(3.48,3.58)	(1.99,2.11)
	40	(.50,.52)	(1.05,1.11)*	(.79,.81)	(3.50,3.56)	(1.94,2.02)
$\frac{1}{2}$	5	(.46,.50)	(.79, .87)*	(.79,.81)	(3.35,3.47)*	(1.74,1.88)*
	10	(.48,.51)	(.91, .99)*	(.80,.82)	(3.39,3.47)*	(1.84,1.96)*
	20	(.49,.51)	(.96,1.02)	(.79,.81)	(3.48,3.54)	(1.94,2.02)
	40	(.49,.51)	(.98,1.02)	(.80,.81)	(3.48,3.52)	(1.96,2.02)
$\frac{1}{4}$	5	(.49,.51)	(.78, .84)*	(.79,.81)	(3.41,3.49)*	(1.73,1.83)*
	10	(.49,.51)	(.87, .93)*	(.80,.81)	(3.44,3.50)	(1.86,1.94)*
	20	(.49,.51)	(.90, .94)*	(.80,.81)	(3.48,3.52)	(1.91,1.97)*
	40	(.50,.51)	(.98,1.01)	(.80,.80)	(3.47,3.51)	(1.95,1.99)*
$\frac{1}{8}$	5	(.43,.45)*	(.73, .78)*	(.78,.80)	(3.31,3.37)*	(1.62,1.68)*
	10	(.48,.50)	(.82, .86)*	(.80,.81)	(3.41,3.45)*	(1.75,1.81)*
	20	(.50,.51)	(.92, .94)*	(.80,.80)	(3.47,3.51)	(1.92,1.96)*
	40	(.50,.50)	(.95, .97)*	(.80,.80)	(3.48,3.50)	(1.95,1.98)*

Table 7. $\hat{Var}(\hat{E}(T_s))$ results from survival distribution simulations. Mean of $\hat{Var}(\hat{E}(T_s))$ is listed for each c and n under each survival distribution.

DISTRIBUTION: c	n	uniform	exponential	beta	normal	gamma
1	5	.0080	.0576	.0037	.1024	.1517
	10	.0044	.0327	.0023	.0531	.0879
	20	.0023	.0193	.0011	.0274	.0489
	40	.0011	.0103	.0006	.0138	.0254
$\frac{1}{2}$	5	.0038	.0250	.0020	.0474	.0710
	10	.0022	.0190	.0011	.0239	.0442
	20	.0011	.0108	.0006	.0135	.0244
	40	.0006	.0056	.0003	.0069	.0130
$\frac{1}{4}$	5	.0017	.0134	.0010	.0195	.0311
	10	.0011	.0087	.0005	.0120	.0210
	20	.0006	.0051	.0003	.0066	.0117
	40	.0003	.0029	.0001	.0034	.0062
$\frac{1}{8}$	5	.0008	.0059	.0005	.0087	.0127
	10	.0005	.0040	.0003	.0057	.0093
	20	.0003	.0025	.0001	.0032	.0057
	40	.0001	.0014	.0001	.0017	.0030

Table 8. $\widehat{\mathrm{Var}}_T(T_s)$ results from survival distribution simulations. Mean of $\widehat{\mathrm{Var}}_T(T_s)$ and (root mean square) are listed for each c and n combination under each survival distribution.

DISTRIBUTION:		uniform	exponential	beta	normal	gamma
$\mathrm{Var}(T_s)$:		.083	1.0	.046	1.0	2.0
c	n					
1	5	.108	.74	.053	1.39	1.90
		(.056)	(.48)	(.027)	(.83)	(1.02)
	10	.108	.84	.062	1.35	2.17
		(.040)	(.47)	(.024)	(.63)	(.80)
	20	.111	1.01	.060	1.34	2.43
		(.034)	(.35)	(.022)	(.46)	(.81)
	40	.105	1.12	.062	1.33	2.54
		(.025)	(.32)	(.018)	(.40)	(.77)
$\frac{1}{2}$	5	.078	.41	.047	.99	1.31
		(.029)	(.66)	(.023)	(.48)	(1.02)
	10	.089	.68	.049	.92	1.69
		(.023)	(.44)	(.015)	(.23)	(.74)
	20	.086	.83	.051	1.06	1.85
		(.015)	(.28)	(.012)	(.22)	(.53)
	40	.088	.88	.053	1.07	2.09
		(.010)	(.21)	(.010)	(.15)	(.38)
$\frac{1}{4}$	5	.069	.35	.048	.72	.91
		(.032)	(.68)	(.017)	(.41)	(1.18)
	10	.080	.50	.045	.90	1.34
		(.014)	(.54)	(.010)	(.28)	(.77)
	20	.083	.64	.047	.97	1.64
		(.013)	(.40)	(.005)	(.16)	(.54)
	40	.086	.85	.047	.98	1.72
		(.008)	(.23)	(.005)	(.11)	(.38)
$\frac{1}{8}$	5	.057	.27	.043	.62	.64
		(.036)	(.75)	(.012)	(.46)	(1.39)
	10	.076	.42	.047	.80	1.08
		(.019)	(.61)	(.008)	(.26)	(1.01)
	20	.085	.62	.046	.91	1.43
		(.009)	(.40)	(.004)	(.15)	(.63)
	40	.083	.75	.047	.96	1.69
		(.006)	(.28)	(.004)	(.09)	(.40)

Table 9. $\hat{\text{Var}}_L(T_s)$ results from survival distribution simulations. Mean of $\hat{\text{Var}}_L(T_s)$ and (root mean square) are listed for each c and n combination under each survival distribution.

DISTRIBUTION:		uniform	exponential	beta	normal	gamma
$\text{Var}(T_s)$:		.083	1.0	.046	1.0	2.0
c	n					
1	5	.095	.58	.045	1.22	1.57
		(.051)	(.59)	(.026)	(.76)	(1.11)
	10	.094	.67	.054	1.18	1.83
		(.034)	(.55)	(.020)	(.56)	(.80)
	20	.097	.84	.053	1.17	2.10
		(.024)	(.39)	(.018)	(.35)	(.69)
	40	.091	.96	.054	1.16	2.21
		(.015)	(.30)	(.012)	(.28)	(.59)
$\frac{1}{2}$	5	.075	.37	.045	.95	1.23
		(.030)	(.70)	(.023)	(.48)	(1.08)
	10	.086	.64	.047	.88	1.60
		(.022)	(.47)	(.014)	(.31)	(.78)
	20	.083	.79	.049	1.02	1.76
		(.015)	(.31)	(.010)	(.22)	(.56)
	40	.084	.84	.051	1.02	2.00
		(.009)	(.24)	(.009)	(.13)	(.36)
$\frac{1}{4}$	5	.068	.34	.048	.71	.89
		(.032)	(.69)	(.017)	(.41)	(1.20)
	10	.079	.49	.044	.89	1.32
		(.014)	(.55)	(.010)	(.29)	(.79)
	20	.083	.63	.046	.96	1.62
		(.013)	(.41)	(.005)	(.16)	(.57)
	40	.085	.84	.046	.97	1.70
		(.008)	(.24)	(.005)	(.11)	(.39)
$\frac{1}{8}$	5	.057	.27	.042	.62	.64
		(.036)	(.75)	(.012)	(.46)	(1.40)
	10	.075	.41	.047	.80	1.07
		(.019)	(.61)	(.008)	(.27)	(1.01)
	20	.085	.62	.046	.91	1.44
		(.010)	(.41)	(.004)	(.15)	(.63)
	40	.083	.75	.047	.96	1.69
		(.006)	(.28)	(.004)	(.09)	(.41)

5.3 Regressions on Estimated Root Mean Squares

We describe the estimated root mean squares ($\hat{RMS}$) of $\hat{E}(T_s)$, $\hat{Var}_T(T_s)$ and $\hat{Var}_L(T_s)$ as a linear function of c and n. Our objective is to determine the response of $\hat{RMS}$ in relation to the changes in Δt and n_i. For each estimator, the regression models were selected based on (1) the form of surface plots of $\hat{RMS}$, c and n, (2) the contribution of significant ($\alpha = .05$) linear, quadratic and crossproduct terms to the full model r^2 using SAS PROC RSREG (Allen, 1982), (3) (a) PROC REG for only linear terms, or (b) backward elimination regressions using PROC STEPWISE (inclusion of term in model at $\alpha = .05$) for models in (2) having quadratic and/or crossproduct terms, and (4) the same (possibly transformed) covariates n and c for each model for each estimator. Models with the following covariates were considered: $\{1/n, c\}$, $\{1/n, \sqrt{c}\}$, $\{1/n, c^2\}$, $\{1/\sqrt{n}, c\}$, $\{1/\sqrt{n}, \sqrt{c}\}$, $\{1/\sqrt{n}, c^2\}$, and $\{1/\sqrt{n}, 1/c\}$.

We determined that the models with covariates 1/n and c satisfactorily met the criteria in (1) to (4) above (Table 10). Surface plots of $\hat{RMS}$'s vs. c and n for $\hat{Var}_T(T_s)$ and $\hat{Var}_L(T_s)$ indicated three groups of survival distributions based on differences in surface shapes. The groups of survival distributions were (1) beta, (2) exponential and gamma, and (3) normal and uniform. Croups were the same for both variance estimators. The differences in surface shapes (groups) are represented by the different regression models (Table 10). From the models, we conclude that, overall, the estimators will better estimate their respective parameter values as the the number of t_i in $[t_0, t_F]$ is increased and as the sample size, n_i, for each t_i is increased.

Table 10. Regression models for $\hat{RMS}$. Model and r^2 for each combination of estimator and survival distribution, $G_s(t)$, are listed. All parameter values are different from zero at $\alpha = .05$.

Estimator	Survival Distribution	Model	r^2
$\hat{E}(T_s)$	beta	$.16(1/n) + .03c$	.89
	exponential	$.02 + 1.29(1/n) + .07c$	.97
	gamma	$.04 + 1.58(1/n) + .10c$	.95
	normal	$1.07(1/n) + .14c$	.94
	uniform	$.35(1/n) + .03c$	.92
$\hat{Var}_T(T_s)$	beta	$.06(1/n) + .02c$	.97
	exponential	$.19 + 5.74(1/n) - .34c - 11.89(1/n\ ) + .36c^2 - 1.89(1/n)c$	.97
	gamma	$.42 + 5.83(1/n) - .73c + 1.03c^2 + 4.49(1/n)c$	.97
	normal	$.08 + 2.02(1/n) - .32c + .62c^2$	.98
	uniform	$.14(1/n) + .02c^2$	.96
$\hat{Var}_L(T_s)$	beta	$.07(1/n) + .01c$	.95
	exponential	$.18 + 5.96(1/n) - .32c - 13.59(1/n^2) + .32c^2 - 1.11(1/n)c$	.98
	gamma	$.34 + 8.17(1/n) - .66c - 11.66(1/n^2) + .77c^2 - 2.72(1/n)c$	.98
	normal	$.03 + 2.14(1/n) + .25c^2$	.96
	uniform	$.15(1/n) + .01c^2$	.94

5.4 Conclusions Based on Simulations

$\hat{E}(T_s)$, $\hat{Var}_T(T_s)$ and $\hat{Var}_L(T_s)$ performed similarly in relation to the shapes of the survival distributions used in the simulations. Based on means and $\hat{RMS}$ of the estimators, these estimators best estimated their respective expected values for the concave beta and uniform survival distributions. These estimators performed worst under the gamma survival distribution. Because in applications the shape of the survival curve is rarely known, possibly graphing the relevant $\hat{p}_{i,s}$ and observing the shape of the graph would aid the researcher in evaluating how 'good' the estimators may be in his/her particular experiment.

$\hat{E}(T_s)$ was a good estimator of $E(T_s)$ for each combination of Δt and n_i, $i = 0,\ldots,F$, for beta, uniform and normal survival distributions. Overall, the larger the number of t_i in $[t_0, t_F]$ and the larger n_i, the closer, on average, $\hat{E}(T_s)$ will estimate $E(T_s)$. Even though $\hat{E}(T_s)$ is a biased estimator of $E(T_s)$ for the exponential and gamma survival distributions, the preceeding recommendation holds because of the decrease in $\hat{RMS}$'s. Also the decrease in $\hat{Var}(\hat{E}(T_s))$'s indicates less variablity of $\hat{E}(T_s)$ as the number of t_i in $[t_0, t_F]$ and n_i are increased. However, in applications, 'large' n_i and 'large' number of

t_i in $[t_0, t_F]$ are difficult to determine unless the researcher has similar information from a previous experiment.

$\hat{\mathrm{Var}}_T(T_s)$ and $\hat{\mathrm{Var}}_L(T_s)$ are reasonable estimators for $\mathrm{Var}(T_s)$ for beta, uniform and normal survival distributions, especially for large n_i and t_i in $[t_0, t_F]$, but rapidly underestimate $\mathrm{Var}(T_s)$ as the number of t_i in $[t_0, t_F]$ increase, especially for smaller n_i. $\hat{\mathrm{Var}}_L(T_s)$ would probably be preferred for a small number of t_i in $[t_0, t_F]$ for concave or straight-line survival distributions (see graphing of $\hat{p}_{i,s}$ above), and $\hat{\mathrm{Var}}_T(T_s)$ would probably be preferred for survival distributions with similar shapes to the normal, exponential or gamma used in the simulations. For large n_i and a large number of t_i in $[t_0, t_F]$ either variance estimator could be used to estimate $\mathrm{Var}(T_s)$ since the two are nearly equal for these conditions as indicated by the simulations and Lemma 1. However, note again that both estimators tend to underestimate $\mathrm{Var}(T_s)$ under exponential and gamma survival distributions.

Regression models of $\hat{\mathrm{RMS}}$ on covariates 1/n and c for each estimator reinforce our previous conclusion that the estimators perform better overall as Δt is decreased and as n_i for each t_i is increased.

6. CONCLUSIONS

Wa recommend the nonparametric estimators outlined in this raport to estimate the respective parameter values under the experiment specified in Section 1. Increasing the number of sample times t_i in $[t_0, t_F]$ and increasing the number of samples, n_i for each t_i will result in better estimates of $E(T_s)$ and $Var(T_s)$. Better estimates of $p_{i,s}$ will be obtained by taking more t_i when frequent stage transitions are occurring. For example, in Section 3, more t_i between $t_3 = 9$ and $t_6 = 15$ would have resulted in better estimates of $p_{i,1}$ and $p_{i,2}$ and, hence, better estimates of $E(T_s)$ and $Var(T_s)$.

Because the estimators appear to best estimate concave or straight-line survival distribution parameter values, graphing $\hat{p}_{i,s}$ may aid the researcher in determining how well the estimators may be performing in his/her experiment. However, based on graphs of $\hat{p}_{i,s}$ from simulations, the graphs of $\hat{p}_{i,s}$ may show considerable variability across t_i and yield no discernable shape of the survival distribution. Because $Var_T(T_s) > Var_L(T_s)$, a conservative overall choice of a variance estimate would be $\hat{Var}_T(T_s)$. However, if a graph of $p_{i,s}$ has a similar shape to one of the survival distributions in Section 5, then the selection of a variance estimate could be based on the simulation results outlined in Section 5.4. If Δt are small and $n_i \geq 20$ then either variance estimate is appropriate. Note again that the experiment requires that the cohort of organisms all begin in stage 0 and complete development in stage A. However, slight deviations from these conditions probably do not significantly effect the estimates, especially for small Δt and large number of n_i.

Some suggested extensions for further research are: (1) using the fact that the $p_{i,s}$ contain doubly censored information to possibly obtain better estimates of parameter values, (2) estimating $E(T_s)$ and $Var(T_s)$ under right and/or left censoring of sample times, and (3) deriving estimators when failures (deaths) can be observed at each t_i.

REFERENCES

Allen, A. (ed). (1982). SAS User's Guide, 1982 Edition, Cary, NC: SAS Institute, Inc.

Bellows, T.S., Jr., Ortiz, M., Owens, J.C., and Huddleston, E.W. (1982). "A Model for Analyzing Insect Stage-Frequency Data When Mortality Varies With Time," Researches on Population Ecology, 24, 142-156.

Birley, M. (1977). "The Estimation of Insect Density and Instar Survivorship Functions From Census Data," Journal of Animal Ecology, 46, 497-510.

Boyer, J.E., Jr. and Deaton, M. (1984). "Estimation of Duration From Stage Frequency Data," Technical Report, Kansas State University, Dept. of Statistics.

Hight, S.C., Eikenbary, R.D., Miller, R.J. and Starks, K.J. (1972). "The Greenbug and Lysiphlebus testaceipes," Environmental Entomology, 1, 205-209.

Kiritani, K., and Nakasuji, F. (1967). "Estimation of the Stage-Specific Survival Rate in the Insect Population With Overlapping Stages," Researches on Population Ecology, 11, 143-152.

Kobayashi, S. (1967). "Estimation of the Individual Number Entering Each Development Stage in an Insect Population," Researches on Population Ecology, 10, 40-44.

Manly, B.F.J. (1974). "Estimation of Stage-Specific Survival Rates and Other Parameters for Insect Populations Developing Through Several Stages," Oecologia, 15, 277-285.

Manly, B.F.J. (1976). "Extensions to Kiritani and Nakasuji's Method for Analysing Insect Stage-Frequency Data," Researches on Population Ecology, 17, 191-199.

Manly, B.F.J. (1977). "A Further Note on Kiritani and Nakasuji's Model for Stage-Frequency Data Including Comments on the Use of Tukey's Jackknife Technique for Estimating Variances," Researches on Population Ecology, 18, 177-186.

Manly, B.F.J. (1985). "Further Improvements to a Method for Analysing Stage-Frequency Data," Researches on Population Ecology, 27, 325-332.

Mills, N.J. (1982). "The Estimation of Mean Duration From Stage Frequency Data," Oecologia, 51, 206-211.

Mood, A.M., Graybill, F.A., and Boes, D.C. (1974). Introduction to the Theory of Statistics, 3rd. ed., New York: McGraw-Hill, Inc.

Ross, H.H. (1965). A Textbook of Entomology, 3rd. ed., New York: John Wiley & Sons.

Wilson, C.L., Loomis, W.E., and Steeves, T.E. (1971). Botany, 5th. ed., New York: Holt, Rinehart & Winston.

APPENDIX A
COMPUTER PROGRAM SOURCE CODE

```
/*SERVICE      UNATTEND
/*REGION       500K
// EXEC SAS,OPTIONS=MACRO
//SYSIN DD *
OPTIONS LS=85;
*@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@;
*@ PROGRAM DURPROG:                                                       @;
*@ ====> LAST REVISED NOVEMBER  06, 1986 BY J.S. PONTIUS.                 @;
*@ CALCULATION OF EXPECTED VALUE, VARIANCE(EXPECTED VALUE) AND            @;
*@ VARIANCE OF TIME TO REACH STAGE S AND DURATION FOR STAGE S FOR ONE @;
*@ SAS DATA SET THAT CONTAINS VALUES OF SAMPLE TIMES AND COUNTS (0, 1,@;
*@ 2, ... ) OF ITEMS IN A STAGE FOR EACH SAMPLE TIME.                     @;
*@ PROGRAM CODED IN SAS MACRO LANGUAGE                                    @;
*@          ( JCL REQUIRED FOR SAS MACRO LANGUAGE:                        @;
*@                   // EXEC SAS,OPTIONS=MACRO                     ).  @;
*@                                                                        @;
*@ NOTE:   DURPROG CAN HANDLE A MAXIMUM OF 20 STAGES.                     @;
*@         (NS = NUMBER OF STAGES).                                       @;
*@ NOTE:   INPUT DATA SET MUST CONTAIN AT LEAST 3 SAMPLE TIMES.           @;
*@                                                                        @;
*@ DURPROG:                                                               @;
*@      (1) READS IN A HARRIS $ADD DATA FILE (CAN SUBSTITUTE DATA CARDS@;
*@          IN PLACE OF $ADD STATEMENT),                                  @;
*@      (2) CHECKS (A) THAT ALL COUNT DATA ARE NONEGATIVE INTEGERS        @;
*@          (B) THAT SAMPLE TIMES ARE >= 0 AND ARE A POSITIVE SEQUENCE.@;
*@      (3) %MACRO _INSECT_:                                              @;
*@          CALCULATES EXPECTED VALUE, VARIANCE(EXPECTED VALUE) AND       @;
*@          VARIANCE OF TIME TO REACH STAGE S AND DURATION FOR STAGE S.@;
*@      (4) %MACRO _LOOPER_: FORMATS OUTPUT FOR PRINTING.                 @;
*@  INPUT DATA SET:                                                       @;
*@          THE FORMAT FOR THE INPUT SAS DATA SET IS AS FOLLOWS           @;
*@          (USER SPECIFIED INFORMATION IN [ ] ):                         @;
*@                                                                        @;
*@          DATA [VALID SAS DATASET NAME];                               *@;
*@            LENGTH DATAID $21;                                         *@;
*@            DATAID= '[STUDY IDENTIFIER OF 1 TO 21 CHARACTERS]';        *@;
*@            INPUT (DAY COUNT1 - COUNT[NS]) (3. [NS]*2.);               *@;
*@            CARDS;                                                     *@;
*@            [DATA ENTERED IN COLUMN FORMAT]                             @;
*@                                                                        @;
*@          REQUIREMENTS: (A) DAY IS THE TIME WHEN SAMPLES WERE TAKEN, @;
*@                        (B) COUNT1 TO COUNT[NS] ARE NONEGATIVE          @;
*@                            INTEGER COUNTS OF ITEMS SAMPLED AT EACH  @;
*@                            SAMPLE TIME.                                @;
*@          EXAMPLE: DATA ONE;                                           *@;
*@                   LENGTH DATAID $21;                                  *@;
```

```
*@                  DATAID= 'STUDY ONE';                             *@;
*@                  INPUT (DAY COUNT1 - COUNT3) (3. 3*2.);           *@;
*@                  CARDS;                                           *@;
*@                   315 0 0                                          @;
*@                  10 9 8 0                                          @;
*@                     .                                              @;
*@                     .                                              @;
*@                     .                                              @;
*@  REFERENCE:                                                        @;
*@                                                                    @;
*@                                                                    @;
*@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@;
*;
*****DEFINE GLOBAL MACRO VARIABLES: NS= NUMBER OF STACES;
*                                   DDS= CURRENT DATA SET;
*                                   STUDYID= DATA SET IDENTIFIER;
*                                   CNT= ARRAY OF COUNTS;
%CLOBAL NS DDS STUDYID CNT ;
*;
************************************************************************;
*  SUBROUTINE _FRMT_:                                                  *;
*  CALLED BY MAIN PROCRAM.                                             *;
*  FORMATS STACES ACCORDING TO VALUE FORMATING FOR PRINTINC RESULTS.   *;
*  INPUTS: F1 - F20 (STAGE IDENTIFIERS),                               *;
*  OUTPUTS: F1 - F20 (FORMATTED STACE IDENTIFIERS).                    *;
************************************************************************
;
*;
%MACRO _FRMT_ (F1,F2,F3,F4,F5,F6,F7,F8,F9,F10,F11,F12,F13,F14,F15,F16,
           F17,F18,F19,F20);
PROC FORMAT ;
      VALUE STAGEFMT 1=&F1
                     2=&F2
                     3=&F3
                     4=&F4
                     5=&F5
                     6=&F6
                     7=&F7
                     8=&F8
                     9=&F9
                    10=&F10
                    11=&F11
                    12=&F12
                    13=&F13
                    14=&F14
                    15=&F15
                    16=&F16
                    17=&F17
                    18=&F18
                    19=&F19
```

```
                20=&F20;
%MEND _FRMT_;
*;
*;
*;
*********************************************************************;
* SUBROUTINE _INSECT_ :                                              *;
* CALLED BY _LOOPER_.                                                *;
* CALCULATION OF EXPECTED DURATION TIMES, VARIANCE OF EXPECTED       *;
* DURATION TIMES, VARIANCE OF DURATION TIMES AND OUTPUT MATRICES FOR*;
* PRINTING RESULTS.                                                  *;
* INPUTS: DAY (UNIT OF TIME MEASUREMENT),                            *;
*         &CNT (ARRAY OF STAGE COUNTS),                              *;
*         INDATA (INPUT DATA SET).                                   *;
* OUTPUTS: OUTDATA (OUTPUT DATA SET OF EXPECTED VALUES & STND DEV OF*;
*                  EXPECTED VALUES).                                 *;
*          OUTDEV  (OUTPUT DATA SET OF STND DEV OF DURATION TIMES.  *;
*********************************************************************;
*;
%MACRO _INSECT_ (DAYVAR,X1,X2,X3,X4,X5,X6,X7,X8,X9,X10,X11,X12,X13,

X14,X15,X16,X17,X18,X19,X20,INDATA=INDTA,OUTDATA=OUTDTA,
               OUTSTD=OUTDEV);
%*;
    PROC MATRIX;
%****************** DAY= VECTOR OF TIMES, COUNT= MATRIX OF STAGE COUNTS;
       FETCH DAY DATA=&INDATA (KEEP=&DAYVAR);
       FETCH COUNT DATA=&INDATA (KEEP=&X1-&&X&NS);
%*;
       N_STAGE=NCOL(COUNT);                    %*NUMBER OF STAGES;
       SAMSIZE=COUNT(,+);                      %*TOTAL SAMPLE SIZE PER;
                                               %*        SAMPLE TIME;
       PROB=COUNT#/(SAMSIZE @ J(1,N_STAGE)); %*MATRIX OF PROPORTIONS;
%*;
%********************IF INITIAL TIME > 0 THEN ADD TIME= 0 TO FOLLOWING;
%*                 MATRICES;
       IF DAY(1,) > 0 THEN DO;
          DAY=J(1,1,0)//DAY;
          WK=J(1,NCOL(PROB),0);
          WK(1,1)=1;
          PROB=WK//PROB;
          SAMSIZE=J(1,1)//SAMSIZE;
          COUNT=J(1,NCOL(COUNT),0)//COUNT;
          COUNT(1,1)=1;
       END;
       N_DAY=NROW(DAY);              %*NUMBER OF SAMPLE TIMES;
%*;
%******************** CALCULATE DIFFERENCE OR SUM BETWEEN I & I+1 TIMES;
                 %* T(I+1) - T(I);
       TIMEINC = DAY(2:N_DAY,)-DAY(1:N_DAY-1,);
```

```
                %* T(I+1)##2 - T(I)##2;
       TIMEINC2= (DAY(2:N_DAY,)##2) - (DAY(1:N_DAY - 1,)##2);
                %* T(I+1) + 2#T(I);
       SM2TERM1= DAY(2:N_DAY,) + (2#DAY(1:N_DAY-1,));
                %* 2#T(I+1) + T(I);
       SM2TERM2= (2#DAY(2:N_DAY,)) + DAY(1:N_DAY-1,);
%********************** GALCULATE DIFFERENCES BETWEEN I-1 & I+1 TIMES;
       T2= TIMEINC(2:N_DAY-1,) + TIMEING(1:N_DAY-2,);
%********************** SQUARE TIME DIFFERENGES FROM PREVIOUS LINE;
       T2= T2#T2;
%*;
%*;
%********************** LOOP THROUGH SUGGESSIVE STACES;
       DO STAGE=2 TO N_STAGE;
          WK=PROB(,1:STAGE-1);
          P=WK(,+);      %*SUM PROPORTIONS IN STACES 1,...,S - 1   ;
%********************** CALCULATE EXPECTED TIME,E(T(S)),TO STACE 2,...,A;
          MU_K_HAT=0.5*(P(1:N_DAY-1,)#TIMEINC + P(2:N_DAY,)#TIMEING);
          MU_K_HAT=MU_K_HAT(+,);
%********************** GALCULATE STANDARD ERROR,SQRT(VAR(E(T)),OF      ;
%*                     EXPECTED TIME TO STAGE 2,...,A                  ;
          STDERR=P#(J(N_DAY,1)-P)#/SAMSIZE; %*BINOMIAL VARIANCE FOR ;
                                            %*     EACH TIME 1,...,F;
          S=0.25*STDERR(2:N_DAY-1,)#T2;     %*VAR FOR TIMES 1 TO F-1;
          S=S(+,);
          STDERR=SQRT(S(1,));
%*;
%********************** CET VEGTOR OF PROPORTIONS OF STACES TO          ;
%*                     CALCULATE EXPEGTED DURATION IN STAGE S* - S,    ;
%*                     E(T(S*) - T(S)).                                ;
          DURAT=PROB(,STACE-1);
%*;
%********************** CALCULATE STANDARD ERROR(EXPECTED DURATION)     ;
%*                     VAR[E(T(S*) - T(S))] FOR STACE S* - S.          ;
%*                     NOTE: FOR STAGE= 2, STDERR= SDIFF.              ;
          SDIFF=DURAT#(J(N_DAY,1)-DURAT)#/SAMSIZE; %*BINOMIAL VAR   ;
                                                  %* FOR EACH TIME.;
          S=0.25*SDIFF(2:N_DAY-1,)#T2;
          S=S(+,);
          SDIFF=SQRT(S(1,));
%********************** GALCULATE SECOND MOMENTS FOR VARIANGES OF TIMES;
%*                     TO REAGH STAGE S, VAR(T(S)).                    ;
          SECMOM1= (P(1:N_DAY-1,) + P(2:N_DAY,))#TIMEINC2;
          SECMOM1= SECMOM1(+,)#/2;
          SECMOM2= P(1:N_DAY-1,)#SM2TERM1 + P(2:N_DAY,)#SM2TERM2;
          SEGMOM2= SECMOM2#TIMEINC;
          SEGMOM2= SECMOM2(+,)#/3;
%*;
%********************** FORMAT RESULTS OF GALCULATIONS FOR PRINTING;
                                   %* IF FIRST PASS THROUCH LOOP ;
```

```
          IF STAGE=2 THEN DO;     %* (IE: STAGE=2)              ;
             DURAT=MU_K_HAT;
             STDIFF=1;
             RDIFF=STDIFF [[ DURAT [[ SDIFF;
             RESULTS=STAGE [[ MU_K_HAT [[ STDERR;
          END;
                                  %* IF STAGE => 2              ;
          ELSE DO;
%********************* CALCULATE EXPECTED DURATION FOR STAGE S* -S;
%*                    E(T(S*) - T(S)), AND FORMAT RESULTS.      ;
           DURAT=MU_K_HAT-RESULTS(STAGE-2,2);
           RDIFF=RDIFF // ((STAGE-1) [[ DURAT [[ SDIFF);
           RESULTS=RESULTS // (STAGE [[ MU_K_HAT [[ STDERR);
         END;
%********************CALCULATE STANDARD DEVIATIONS OF TIME TO REACH  ;
%*                  STAGE S, SQRT[VAR(T(S))].                       ;
         VAR1= SECMOM1 - (MU_K_HAT##2);
         STDEV1= SQRT(VAR1);
         VAR2= SECMOM2 - (MU_K_HAT##2);
         STDEV2= SQRT(VAR2);
%*                  FORMAT STANDARD DEVIATION MATRICES FOR OUTPUT;
         IF STAGE= 2 THEN STDRES= STAGE [[ STDEV1 [[ STDEV2;
         ELSE DO;
           TEMP= STAGE [[ STDEV1 [[ STDEV2;
           STDRES= STDRES // TEMP;
         END;
       END;
%************************* FORMAT MATRICES FOR OUTPUT;
       S=1;  M=0;  V=0;
       RESULTS=(S [[ M [[ V) // RESULTS;
       S=N_STAGE;
       RDIFF=RDIFF // (S [[ M [[ V);
       RESULTS=RESULTS [[ RDIFF(,2:3);
       OUTPUT RESULTS OUT=&OUTDATA (RENAME=(COL1=STAGE COL2=ESTIMATE
                                     COL3=STDERR COL4=DIFF
                                     COL5= STD_DIFF));
       OUTPUT STDRES  OUT=&OUTSTD  (RENAME=(COL1= STAGE COL2= SD1
                                     COL3= SD2));
%MEND _INSECT_;
*;
*;
*;
*************************************************************************;
* SUBROUTINE _LOOPER_:                                                  *;
*  CALLED BY MAIN PROGRAM.                                              *;
*  PRINTS INPUT DATA SET, CALLS _INSECT_ FOR DATA PROCESSING, CALLS     *;
*  PROC MEANS FOR CALCULATION OF USUAL MEANS AND STANDARD DEVIATIONS    *;
*  OF DURATION TIMES, AND FORMATS RESULTS FOR OUTPUT PRINTING.          *;
*  INPUTS: DDS (INPUT DATA SET),                                        *;
*          UNIT   (UNIT OF SAMPLE TIME MEASUREMENT).                    *;
```

```
* OUTPUTS: NONE.                                                       *;
*************************************************************************;
*;
%MACRO _LOOPER_(INDSET,UNIT);
        DATA DATA1;
           SET &INDSET;
           CALL SYMPUT('STUDYID',DATAID);
%****************** PRINT OUTPUT HEADER AND DATA SET;
        PROC PRINT;
           TITLE1 'ESTIMATION OF TIME TO AND DURATION OF STAGE';
           TITLE2 'FREQUENCY DATA, METHOD OF BOYER AND DEATON.';
           TITLE3 'PROGRAM REVISED: SEPTEMBER, 1986 BY JS PONTIUS.';
           TITLE4 STUDY: &STUDYID;
           TITLE5 UNIT OF TIME MEASUREMENT: &UNIT;
           VAR DAY COUNT1-COUNT&NS;
           %_INSECT_(DAY,&CNT,
                    INDATA=%SCAN(&SYSDSN,2),OUTDATA=DTA1,OUTSTD=DTA2);
%***************** SET EXPECTED, VARIANCE OF AND VARIANCE OF EXPECTED ;
%*                 DURATION TIMES TO MISSING FOR FIRST AND LAST STAGES;
        DATA DTA1;
           SET DTA1;
           IF STAGE=1 THEN DO;
              ESTIMATE=.;
              STDERR=.;
           END;
           IF STAGE=&NS THEN DO;
              DIFF=.;
              STD_DIFF=.;
           END;
%********************* PRINT RESULTS OF EXPECTED VALUES & VAR(EXPECTED;
%*                     VALUES).                                        ;
        PROC PRINT SPLIT='#' DATA= DTA1;
           ID STAGE;
           VAR ESTIMATE STDERR DIFF STD_DIFF;
           FORMAT STAGE STAGEFMT.;
           LABEL ESTIMATE='  TIME TO#REACH STAGE#( E(T(S)) )'
                 DIFF='  DURATION TIME#( E(T(S)) - T(S'')) )'
                 STDERR='STD ERROR OF#  E(T(S))'
                 STD_DIFF='  STD ERROR OF#E(T(S) - T(S''))';
%********************** PRINT RESULTS OF VAR(T(S));
        PROC PRINT SPLIT='#' DATA= DTA2;
          ID STAGE;
          VAR SD1 SD2;
          FORMAT STAGE STAGEFMT.;
          LABEL SD1= 'STD DEVIATION OF T(S)#-TRAPEZOID ANALOG-'
                SD2= 'STD DEVIATION OF T(S)# -STRAIGHT LINE-';
%MEND _LOOPER_;
*;
*;
*;
```

```
*********************************************************************;
* SUBROUTINE _DATACK_:                                              *;
*   CALLED BY MAIN PROGRAM.                                         *;
*   CHECKS INPUT DATA SET FOR NEGATIVE AND IMPROPERLY SEQUENCED     *;
*   SAMPLE TIMES.                                                   *;
*   CHECKS INPUT DATA SET FOR ILLEGAL NEGATIVE AND REAL COUNT DATA. *;
*   INPUTS: &DDS  (INPUT DATA SET).                                 *;
*   OUTPUTS: NONE.                                                  *;
*********************************************************************;
%MACRO _DATACK_;
  DATA CHECK;
    SET &DDS;
%************ CHECK FOR SAMPLE TIME(J) <= SAMPLE TIME(J - 1);
  PROC MATRIX;
    FETCH OBSDATA DATA= CHECK;
    STIME= OBSDATA(,1);
    NDAY= NROW(STIME);
    IF NDAY => 2 THEN DO;
      OUTMAT= J(NDAY,2,0);
      DO K= 2 TO NDAY;
        IF ABS(STIME(K,)) <= ABS(STIME((K-1),)) THEN DO;
          OUTMAT(K,1)= K;
          OUTMAT(K,2)= STIME(K,);
        END;
      END;
    END;
    OUTPUT OUTMAT OUT= DAYERR (RENAME= (COL1= OBSNUMB COL2= DAY));
  DATA DAYERR;
    SET DAYERR;
    IF OBSNUMB > 0 THEN DO;
      PUT '----->ERROR: SAMPLE TIME IS LESS THAN PREVIOUS SAMPLE TIME.';
      PUT '              ' OBSNUMB=;
      PUT '              ' DAY=;
    END;
  DATA CHECK;
    SET CHECK;
%*************** CHECK FOR SAMPLE TIME < 0.0;
    IF DAY < 0.0 THEN DO;
      PUT '----->ERROR: SAMPLE TIME < ZERO.';
      PUT '             ' _N_= ;
      PUT '             ' DAY=;
    END;
%*;
    %DO I= 1 %TO &NS;
%*************** CHECK FOR NEGATIVE COUNT DATA;
      IF COUNT&I  < 0.0 THEN DO;
        PUT '----->ERROR: COUNT DATA VALUE IS NEGATIVE.';
        PUT '             ' _N_=;
        PUT '             ' COUNT&I=;
      END;
```

```
%*************** CHECK FOR REAL COUNT DATA;
      IF (COUNT&I - INT(COUNT&I)) > 0.0 THEN DO;
        PUT '------>ERROR: COUNT DATA VALUE IS NOT AN INTEGER.';
        PUT '              ' _N_=;
        PUT '              ' COUNT&I=;
      END;
    %END;
  PROC DELETE DATA= CHECK DAYERR;
%MEND _DATACK_;
*;
*;
*;
*;
*************************************************************************;
* MAIN PROGRAM:                                                          *;
* STATEMENTS REQUIRED FOR PROGRAM EXECUTION:                             *;
*      (1) HARRIS $ADD DATA FILE (OR SUBSTITUTE DATA SET AS DESCRIBED    *;
*          IN PROGRAM HEADER),                                           *;
*      (2) SPECIFY THE NUMBER OF STAGES TO BE ANALYZED (IE: NS)          *;
*      (3) ASSIGN THE ARRAY, CNT, ELEMENTS OF VARIABLES COUNT1 TO        *;
*          COUNT[NS],                                                    *;
*      (4) ENTER LABELS FOR STAGES IN _FRMT_ SUBROUTINE PARAMETER LIST,*;
*      (5) ENTER SAMPLE TIME UNIT (MEASUREMENT) IN 2ND SLOT OF           *;
*          PARAMETER LIST IN SUBROUTINE _LOOPER_.                        *;
*************************************************************************;
*;
*****ADD DATA CARDS HERE:     ******************************************;
*(1);
$ADD LTESTA
*****MAIN PROGRAM STATEMENTS BEGIN HERE:     ***************************;
              %LET DDS=%SCAN(&SYSDSN,2);
*(2);         %LET NS= 3;
              %PUT NOTE: DATA SETS CHECK & DAYERR ARE FOR ERROR
                         ROUTINES.;
              %_DATACK_;
*(3);         %LET CNT=COUNT1,COUNT2,COUNT3;
              %PUT NOTE: DATA SET CHECK IS FOR ERROR ROUTINES.;
              %PUT NOTE: DATA SETS CHECK & DAYERR ARE BEING DELETED.;
*(4);         %_FRMT_(EGGLARVA,PUPA,ADULT);
*(5);         %_LOOPER_(&DDS,DAYS);
/*
```

NONPARAMETRIC ESTIMATION OF STAGE TRANSITION TIME
FROM STAGE FREQUENCY DATA
by

Jeffrey S. Pontius

B.A., Millersville University, 1976
M.S., North Dakota State University, 1982

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1987

ABSTRACT: We derive and evaluate, by simulation, estimators for the following experiment. Consider an organism that displays observable stages. Choose a sequence of fixed sample points in time. At each sample point, observe a subset of a cohort of organisms and record the number of organisms in each stage. Our objective is to estimate parameters of the time in stage s, T_s, for an organism. We review estimators of the time to stage s, $\hat{E}(T_s)$, mean duration time, $\hat{E}(T_s - T_{s'})$, and the variance of $\hat{E}(T_s)$, $\hat{Var}(\hat{E}(T_s))$, proposed by Boyer and Deaton (1984). We derive two variance estimators and prove two relational properties. Simulation results under five survival distributions indicate that the estimators provide reasonable estimates of parameter values. The estimators better estimate the parameter values as the number of sample times is increased in a finite interval and as the number of samples per sample time is increased. The estimators are useful in studies on survival data, quality control, and other studies in life sciences and engineering. We also describe a computer program to calculate the estimates.