NAVAL RESEARCH LOGISTICS QUARTERLY

DECEMBER 1982
VOL. 29, NO. 4

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $15.00 a year in the U.S. and Canada, $18.75 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

The Naval Research Logistics Quarterly is published from appropriated funds by authority of the Office of Naval Research in accordance with NPPR P-35. Controlled circulation postage paid at Washington and additional mailing offices. Articles, letters and address changes may be forwarded to Department of the Navy, Office of Naval Research, Ballston Tower # 1. Room 607, 800 N. Quincy St., Arlington, VA. 22217.

PERIODIC REPLACEMENT WHEN MINIMAL REPAIR COSTS VARY WITH TIME*

Philip J. Boland

Department of Mathematics
University College Dublin
Belfield, Dublin, Ireland

ABSTRACT

A policy of periodic replacement with minimal repair at failure is considered for a complex system. Under such a policy the system is replaced at multiples of some period T while minimal repair is performed at any intervening system failures. The cost of a minimal repair to the system is assumed to be a nondecreasing function of its age. A simple expression is derived for the expected minimal repair cost in an interval in terms of the cost function and the failure rate of the system. Necessary and sufficient conditions for the existence of an optimal replacement interval are exhibited in the case where the system life distribution is strictly increasing failure rate (IFR).

1. INTRODUCTION

A complex system may fail if one of its many components ceases to function. The system is returned to the operating state when the failed component is replaced. As the great majority of components have not been replaced, the remaining life distribution and failure rate of the system are essentially undisturbed. This type of system repair, whereby the failure rate of the system is not altered by the failure and subsequent repair of the system is known as minimal repair.

A policy of periodic replacement with minimal repair at failure is one in which the system is replaced at multiples of some period T while performing minimal repair at any intervening system failures. This type of policy was introduced and investigated by Barlow and Hunter [1] (see also Barlow and Proschan [2]). In their paper, Barlow and Hunter show how to calculate the optimal period T assuming the cost of a minimal repair is constant and using as an optimality criterion the minimization of total expected cost per unit time over an infinite time horizon. Tilquin and Cleroux [8] modify this model by introducing to the cost analysis a general cost function (to account for depreciation or adjustment costs, interest charges, monitoring costs and the like), which increases continuously with the length of time the system is in use. However, as in the model considered by Barlow and Hunter, the cost of a minimal repair to the system is constant (and hence, for example, does not depend on the time at which the repair is made or the previous number of repairs to the system). Boland and Proschan [3] generalize the Barlow-Hunter model to incorporate the situation when the cost of a minimal repair is an increasing function of the number of previous repairs to the system. Further generalizations of the Barlow and Hunter model can be found in the work of Nakagawa (see for example Nakagawa [5]).

*Research supported by Air Force Office of Scientific Research, U.S.A.F., under Contract No. F49620-82-K-0007.

In this paper, it is assumed that the cost of a minimal repair to the system which fails at age t is $C(t)$, where $C(t)$ is a continuous nondecreasing function of t. Hence, as the system ages it becomes more expensive to perform minimal repair. F will denote the life distribution function of the system with density f. $r(t)$ will denote the failure rate function of the system, while $R(t) = \int_0^t r(s)ds$ is the cumulative failure rate function or hazard function of the system. It is assumed that $F(0) = 0$ and that $r(t)$ is a positive continuous function of t. The problem of finding an optimal period T_0 for periodic replacement (where minimal repair is performed according to the cost function $C(t)$ on system failure) is investigated. It is shown that if $C(t)r(t)$ is nondecreasing (in particular if F is IFR), then an optimal period T_0 (possibly infinite) exists.

2. EXPECTED MINIMAL REPAIR COSTS

Now consider the situation where the system is in operation in the time interval $[0, T)$. On failure, minimal repair is made (according to $C(t)$) and we assume that repair time is negligible. If the system is not replaced in this interval, the following theorem gives two useful expressions for the expected costs of minimal repair in $[0, T)$.

THEOREM 1: The expected minimal repair cost of the system in the interval $[0, T)$ is

$$\int_0^{R(T)} C(R^{-1}(t))dt = \int_0^T C(t)r(t)dt.$$

PROOF: Let N_T be the random variable denoting the number of minimal repairs performed on the system in the age interval $[0, T)$. We know that N_T has a Poisson distribution with parameter $R(T)$ (see Barlow and Proschan [2], pp. 96-97 and Cox and Miller [4], p. 153).

Now if $N_T = k$, and $t_1, \ldots, t_k$ are the times of the minimal repairs, then the total minimal repair cost in the interval $[0, T)$ is $\sum_{i=1}^{k} C(t_i)$. Given $N_T = k$, we know that $\tau_1 = R(t_1), \ldots, \tau_k = R(t_k)$ are distributed as the order statistics of a random sample of size k from the uniform distribution on $[0, R(T))$ (see, for example, Parzen [6], pp. 139-143 or Thompson [7]). Hence, the expected minimal repair cost given $N_T = k$ is

$$\begin{aligned}
&E(C(t_1) + \ldots + C(t_k)|N_T = k) \\
&\quad = E(C(R^{-1}(\tau_1) + \ldots + C(R^{-1}(\tau_k))|N_T = k) \\
&\quad = kE(C(R^{-1}(\tau))|N_T = k) \\
&\qquad\qquad \text{(where } \tau \text{ is uniformly distributed on } [0, R(T))) \\
&\quad = k\int_0^{R(T)} C(R^{-1}(t))\,\frac{1}{R(T)}\,dt \\
&\quad = \frac{k}{R(T)}\int_0^{R(T)} C(R^{-1}(t))dt.
\end{aligned}$$

Therefore, the expected minimal repair cost in the interval $[0, T)$ is

$$E_{N_T}(E(C(t_1) + \ldots + C(t_k)|N_T = k))$$

$$= E_{N_T}\left[\frac{k}{R(T)}\int_0^{R(T)} C(R^{-1}(t))\,dt\right]$$

$$= \left[\frac{1}{R(T)}\int_0^{R(T)} C(R^{-1}(t))\,dt\right][E_{N_T}(k)]$$

$$= \left[\frac{1}{R(T)}\int_0^{R(T)} C(R^{-1}(t))\,dt\right][R(T)]$$

$$= \int_0^{R(T)} C(R^{-1}(t))\,dt$$

$$= \int_0^T C(w)r(w)\,dw$$

(using the change of variable $w = R^{-1}(t)$).

REMARK 1: In the above proof of Theorem 1, the number of minimal repairs to the system in $[0, T)$ is a nonhomogeneous (or homogeneous) Poisson process with intensity function $r(t)$. In particular, given that the system is functioning at age t, the probability of a system failure in the interval $(t, t + h)$ is of the form $r(t)h + o(h)$. Moreover, the mean number of failures in the interval $[0, T)$ is $\int_0^T r(t)\,dt = R(T)$. This leads us to interpret $C(t)r(t)$ in a naive sort of way as the "rate" of spending a dollar on minimal repair at age t. With this interpretation $\int_0^T C(t)r(t)\,dt$ represents the mean number of dollars spent on minimal repair in $[0, T)$, which we have rigorously demonstrated in Theorem 1.

We now consider some examples of minimal repair cost functions and determine the resulting expected minimal repair cost in an age interval $[0, T)$ for a given life distribution. It is reasonable to assume that the cost of a minimal repair to a system should be a nondecreasing function of its age. Perhaps even more appropriate would be a nondecreasing function of the life distribution $F(t)$ or equivalently $R(t)$ or $\frac{1}{\bar{F}(t)}$.

EXAMPLE 1: Let $C(t)$ be of the form $C(t) = g(R(t))$ where g is nondecreasing and $G' = g$. Then

$$\int_0^T C(t)r(t)\,dt = \int_0^T g(R(t))\,dR(t) = G(R(t)) - G(0).$$

We consider the following particular cases:

(a) $g(y) = cy^\alpha$ $(\alpha \geqslant 0)$. Then $C(t) = c\left[\log\frac{1}{\bar{F}(t)}\right]^\alpha$ and $\int_0^T C(t)r(t)\,dt = c\,\frac{R^{\alpha+1}(T)}{\alpha+1}$. When $\alpha = 1$, i.e., minimal repair cost is proportional to the cumulative failure rate function, then the expected minimal repair cost in $[0, T)$ is

$c\,\frac{R^2(T)}{2}$. When $\alpha = 0$, one obtains the cost function used by Barlow and Hunter [1].

(b) $g(y) = ce^{\alpha y}$ $(\alpha > 0)$. Then $C(t) = g(R(t)) = ce^{\alpha R(t)} = c\,\frac{1}{(\bar{F}(t))^{\alpha}}$ and $\int_0^T C(t)r(t)\,dt = \frac{c}{\alpha}\,(e^{\alpha R(T)} - 1) = \frac{c}{\alpha}\left[\frac{1}{\bar{F}^{\alpha}(T)} - 1\right]$. When $\alpha = 1$, i.e., the cost of a minimal repair is inversely proportional to the survival probability, then the expected minimal repair cost in $[0, T)$ is proportional to the odds ratio $F(T)/\bar{F}(T)$.

The linear combinations of (a) and (b) constitute a large class of reasonable cost functions for minimal repair.

EXAMPLE 2: One might consider a cost function of the type $C(t) = ct^{\alpha}$ $(\alpha \geqslant 0)$. In the particular case when $\alpha = 1$ this yields an expected minimal repair cost in $[0, T)$ of

$$\int_0^T ctr(t)\,dt = cR(T)T - \int_0^T cR(t)\,dt.$$

If the life distribution is Weibull of the form $F(t) = 1 - e^{-(\lambda t)^{\beta}}$, then the expected cost is

$$c(\lambda T)^{\beta}T - \int_0^T c(\lambda t)^{\beta}\,dt = c\lambda^{\beta}T^{\beta+1}\left[\frac{\beta}{\beta+1}\right].$$

3. PERIODIC REPLACEMENT WITH MINIMAL REPAIR

Let us now consider the problem of finding a period T_0 for replacement which minimizes expected long-run cost per unit of time. c_0 will denote the cost of a planned system replacement. If $\mathscr{C}(T)$ represents the expected long-run cost per unit of time when the system is periodically replaced at times $T, 2T, 3T, \ldots$, then $\mathscr{C}(T)$ has the form

$$\mathscr{C}(T) = \frac{\int_0^T C(t)r(t)\,dt + c_0}{T}$$

Therefore,

$$\mathscr{C}'(T) = \frac{C(T)r(T)T - \int_0^T C(t)r(t)\,dt - c_0}{T^2},$$

and this yields

THEOREM 2: If $C(t)r(t)$ is a nondecreasing function of t, then an optimal replacement interval T_0 exists. T_0 is finite if

$$\lim_{T\to\infty} \int_0^T [C(T)r(T) - C(t)r(t)]\,dt > c_0$$

or if $C(t)r(t)$ is eventually constant and

$$\lim_{T\to\infty} \int_0^T [C(T)r(T) - C(t)r(t)]dt \geqslant c_0,$$

otherwise the optimal policy is to never replace.

PROOF: If $C(t)r(t)$ is nondecreasing then $C(T)r(T)T - \int_0^T C(t)r(t)\,dt$ is nondecreasing and hence $C'(T)$ has a zero iff

$$\lim_{T\to\infty} \int_0^T [C(T)r(T) - C(t)r(t)]dt > c_0$$

or

$$\lim_{T\to\infty} \int_0^T [C(T)r(T) - C(t)r(t)]dt \geqslant c_0$$

if $C(T)r(T)$ is eventually constant.

REMARK 2: If F is an IFR distribution with strictly increasing failure rate function $r(t)$, then there exists a unique optimal replacement interval T_0 (possibly infinite).

EXAMPLE 3: Let us consider $C(t)$ of the form $C(t) = ce^{\alpha R(t)}$ (see Example 1(b)): if F is Weibull of the form $F(t) = 1 - e^{-(\lambda t)^\beta}$ where $\beta > 1$, then the optimal period T_0 must satisfy

$$ce^{\alpha(\lambda T_0)^\beta}\left[\beta\lambda^\beta T_0^\beta - \frac{1}{\alpha}\right] = c_0 - \frac{c}{\alpha}.$$

If $\alpha = 0$, then $T_0 = \frac{1}{\lambda}\left(\frac{c_0}{(\beta-1)c}\right)^{1/\beta}$.

EXAMPLE 4: Let $C(t)$ be of the form $C(t) = c(R(t))^\alpha$ and F be Weibull where $F(t) = 1 - e^{-(\lambda t)^\beta}$. Then $C(t)r(t) = c\beta\lambda(\lambda t)^{\alpha\beta+\beta-1}$. This is increasing iff $\alpha\beta + \beta > 1$, and hence, in particular, if $\beta > 1$ (or for example $\beta = 1/2$ (F is DFR) and $\alpha = 3/2$) there is a unique optimal replacement interval T_0.

REMARK 3: We may also wish to consider the problem of finding a replacement interval T_0 which minimizes total expected costs over a finite time horizon $[0, t_0)$. For period T, where $mT < t_0 < (m+1)T$ for some integer m, we see that the total expected cost in $[0, t_0)$ is

$$\mathscr{C}_{t_0}(T) = mc_0 + m\int_0^T C(t)r(t)\,dt + \int_0^{t_0-mT} C(t)r(t)\,dt.$$

If $C(t)r(t)$ is an increasing function of t, it follows that $\mathscr{C}_{t_0}'(T) = m[C(T)r(T) - C(t_0 - mT)r(t_0 - mT)] \geqslant 0$ in $\left(\frac{t_0}{m+1}, \frac{t_0}{m}\right]$. Therefore, if $C(t)r(t)$ is increasing then the optimal replacement interval T_0 is an element in the set $\{t_0, t_0/2, t_0/3, \ldots\}$.

REFERENCES

[1] Barlow, R.E. and L.C. Hunter, "Optimal Preventative Maintenance Policies," Operations Research *8*, 90-100 (1960).

[2] Barlow, R.E. and F. Proschan, *Mathematical Theory of Reliability* (John Wiley and Sons, New York, 1965).

[3] Boland, P.J. and F. Proschan, "Periodic Replacement with Increasing Minimal Repair Costs at Failure," Operations Research *30*, 1183-1189 (1982).

[4] Cox, D.R. and H. Miller, *The Theory of Stochastic Processes* (Chapman and Hall, London, 1965).

[5] Nakagawa, T., "A Summary of Periodic Replacement with Minimal Repair at Failure," Journal of Operations Research Society of Japan *24*, 3, 213-228 (1981).

[6] Parzen, E., *Stochastic Processes,* Holden Day Series in Probability and Statistics (Holden-Day, Inc. San Francisco, CA, 1962).

[7] Thompson, W.A., Jr., "On the Foundations of Reliability," Technometrics *23*, 1, 1-13 (1981).

[8] Tilquin, C. and R. Cleroux, "Period Replacement with Minimal Repair at Failure and General Cost Function," Journal of Statistical Computing and Simulation *4*, 63-67 (1975).

RELIABILITY OF A SINGLE UNIT WITH K SPARES IN CASE OF THE HYPER-EXPONENTIAL AND THE ERLANG FAILURE DISTRIBUTION

J. van Dam and F. J. M. Raaijmakers

Royal Netherlands Naval College
Den Helder, Netherlands

ABSTRACT

In this paper formulas are derived for the reliability of a single unit to which identical spares in standby are allocated, with all of these units having a hyper-exponential or Erlang distributed lifetime. Two advantages are obtained by using these distributions. First, the general procedure for calculating the effect of redundancy is applicable, in contrast with most of the common life distributions, such as Weibull. Moreover, both distributions are suitable for matching to observed curves and practical data by varying the values of the parameters, covering together most of the cases of practical interest.

1. INTRODUCTION

A complex system made up of several subsystems in series, cannot successfully perform its function, unless every subsystem remains in operation. In case of malfunctioning of one of them, a spare must be present in order to replace the failing part on-line. If the needed spare part is not available the series-system is out of service, while the remaining spares for the other subsystems are of little use for the moment.

To increase the reliability of temporarily isolated systems—e.g., aboard a ship on route—spare parts must be taken along in a sufficient number. A shortage during this period of isolation not only decreases the availability of the system concerned but also entails an enormous rise in the costs of maintenance, if this component has to be supplied on an emergency basis, if possible at all.

In former days it was possible to provide the system with an abundant stock of spares. If there were no budgetary, spatial or other limitations the stocks taken along could be extended to such a high level that for a certain period there was practically no shortage of any spare. Nowadays, however, due to increased complexity and consequently high prices of modern subsystems, the costs of such a stock would be prohibitive.

A procedure has been developed based on dynamic programming by means of which the composition of an optimal kit can be calculated [9]. The basic idea is that the chance of a shortage of any spare and a simultaneous surplus of the other ones is as small as possible. An optimal kit, of all kits of the same cost, guarantees the highest reliability of the series-system to which the kit is joined.

The spare part provisioning in behalf of modern systems is a logistics problem, which cannot be solved satisfactorily only by best engineering judgement based upon intuition and experience, as happened formerly. A numerical approach must be made at present with the aid of the theory of reliability, using formulas to determine the chance of survival of a unit to which one or more units in standby are allocated.

Many papers in this field of investigation assume an exponentially distributed lifetime of the units and spare parts concerned. This distribution is a one parameter distribution and completely defined when the mean lifetime is specified. Characteristic is the constant failure rate which means that the probability of failure is independent of age: the unit is as good as new at any moment of its lifetime. A certain justification for assuming an exponentially distributed lifetime is given by Palm's theorem [5] and, of course, by so many units actually showing an exponentially distributed lifetime. On the other hand the exponential distribution is not flexible enough to describe the survival curve of many other components or systems. Therefore, other failure distributions (such as Weibull, gamma, etc.) are proposed [7,3,4]. These distributions are characterized by more than one parameter and so are more suited for fitting to observed reliability curves in practice.

A drawback of most of the nonexponential distributions is that it is difficult, if not impossible, to obtain a closed form expression for the effect of redundancy by using the general procedure to derive an expression for the reliability of a unit with one or more spares. In this paper two generalized exponential distributions are studied, together covering most of the cases of practical interest; for both distributions a formula for the reliability with spares is derived.

2. RELIABILITY OF A SINGLE UNIT WITH K SPARES

In this section the reliability of a single unit together with k (identical) spares is derived. The procedure holds for every life distribution [2,6].

Figure 1 shows a single unit (equipment, component, system) on-line, to which k identical spares are allocated, $k = 0, 1, 2, \ldots$. When unit 1 fails, unit 2 in standby is put in operation, while unit 1 is taken off-line. This process will be repeated until the kit of k spares is exhausted.

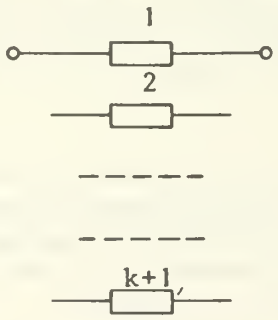


FIGURE 1. Unit 1 on-line, k spares allocated.

Let X_i be the lifetime of the ith unit, $i = 1, \ldots, k + 1$. Then the lifetime Y_{k+1} of the whole system consisting of one unit with k spares may be written as

$$Y_{k+1} = \sum_{i=1}^{k+1} X_i.$$

Let $a(t)$ be the same probability density function (p.d.f.) for all X_i, $f(t)$ the p.d.f. of Y_{k+1}. Then renewal theory states the following relationship between the two functions f and a:

$$Lf = (La)^{k+1},$$

where Lg is the Laplace transform of a function g. Using the inverse transformation one finds the function f.

Finally, we can find $R_{k+1}(t)$ by integration of $f(t)$:

$$R_{k+1}(t) = \int_t^\infty f(u)\,du, \quad t \geqslant 0. \tag{2.1}$$

This general procedure, if applied to the exponential distribution $a(t) = \lambda e^{-\lambda t}$, gives the well-known results:

$$f(t) = \frac{\lambda^{k+1}}{k!}\, t^k e^{-\lambda t}$$

$$k = 0, 1, \ldots, \tag{2.2}$$

$$R_{k+1}(t) = e^{-\lambda t} \sum_{i=0}^{k} \frac{(\lambda t)^i}{i!}.$$

3. THE HYPER-EXPONENTIAL AND ERLANG DISTRIBUTIONS

As mentioned in the introduction, most of the common life distributions unfortunately are unsuitable for the application of the general procedure described in Section 2. In most cases it is difficult or even impossible to calculate the Laplace transform of the density function $a(t)$. Even if the transform of $a(t)$ can be calculated, the expression is so complex that calculation of the inverse transform is difficult. Although in many cases numerical methods are available for common distributions, such as the Weibull, closed form expressions for the effect of spare parts in standby are not yet known. It must be emphasized that this lack of knowledge of a closed form expression must be seen as a more and more disturbing drawback, as the redundancy is of increasing importance for an optimal spare part policy. Now it seems that two distributions are available which combine the two operational advantages, namely, the suitability both for curve fitting and for application of the procedure in Section 2. These two distributions expressed by their reliability functions are, respectively:

1. The hyper-exponential distribution:

$$R_1(t) = \omega e^{-\delta\mu t} + (1-\omega)e^{-(1-\delta)\mu t} \qquad \begin{matrix} 0 \leqslant \omega \leqslant 1 \\ 0 < \delta < 1 \\ \mu > 0 \end{matrix} \tag{3.1}$$

2. The Erlang distribution:

$$R_1(t) = e^{-L\lambda t} \sum_{i=0}^{L-1} \frac{(L\lambda t)^i}{i!} \qquad L = 1, 2, \ldots, \quad \lambda > 0. \tag{3.2}$$

Both distributions can be seen as generalizations of the exponential distribution: the hyper-exponential is exponential if $\omega = 0$, $\omega = 1$ or $\delta = 1/2$ while the Erlang is exponential in case $L = 1$.

To compare these distributions we first give the results for mean and variance. θ_u is the mean of the hyper-exponential distribution.

	Mean	Variance
Hyper-exponential	$\left[\frac{\omega}{\delta} + \frac{1-\omega}{1-\delta}\right]\frac{1}{\mu}$	$\left[1 + \frac{2\omega(1-\omega)(1-2\delta)^2}{(\omega - 2\omega\delta + \delta)^2}\right] \cdot \theta_u^2$
Erlang	$\frac{1}{\lambda}$	$\frac{1}{L\lambda^2}$

Characteristic is that the standard deviation is always less than or equal to the mean in case of an Erlang distribution where the opposite is true for the hyper-exponential distribution.

The hazard function of the hyper-exponential distribution is given by

$$z(t) = \mu \frac{\delta\omega + (1-\delta)(1-\omega)e^{-(1-2\delta)\mu t}}{\omega + (1-\omega)e^{-(1-2\delta)\mu t}}. \tag{3.3}$$

Simple analysis gives that z is a monotonic decreasing function and

$$\lim_{t\to\infty} z(t) = \delta\mu.$$

In Figure 2 some graphs of the hazard function are given for particular values of the parameters (resulting in the same mean value).

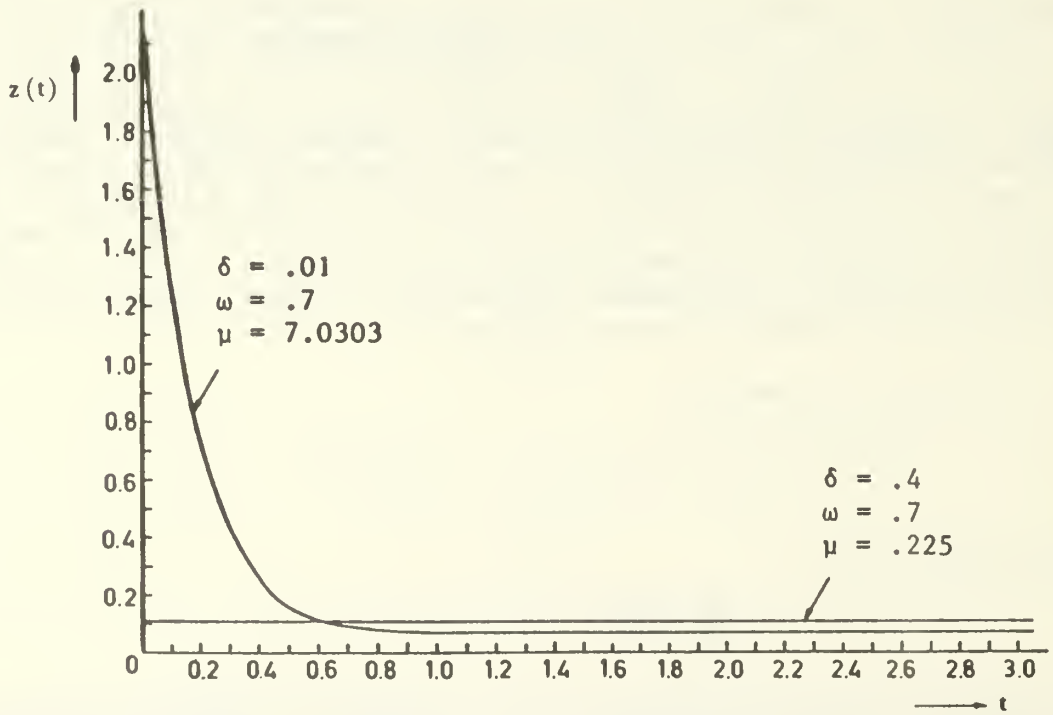


FIGURE 2. Graphs of the hazard function for the hyperexponential distribution.

With the hyper-exponential distribution we are able to describe a high failure rate in the first part of the lifetime of a unit. After a "burn-in" period the failure is approximately constant.

The hazard function of the Erlang distribution is an increasing function of time [1] and $\lim_{t\to\infty} z(t) = L\lambda$ in this case. So those two distributions form a group whose standard deviations cover all values from zero to infinity and the same is true for their hazard functions. These distributions have proven to represent adequately many real lifetimes of systems and units [6]. Figure 3 shows the reliability function of both Erlang and hyper-exponential distributions, where time is normalized with respect to the mean lifetime.

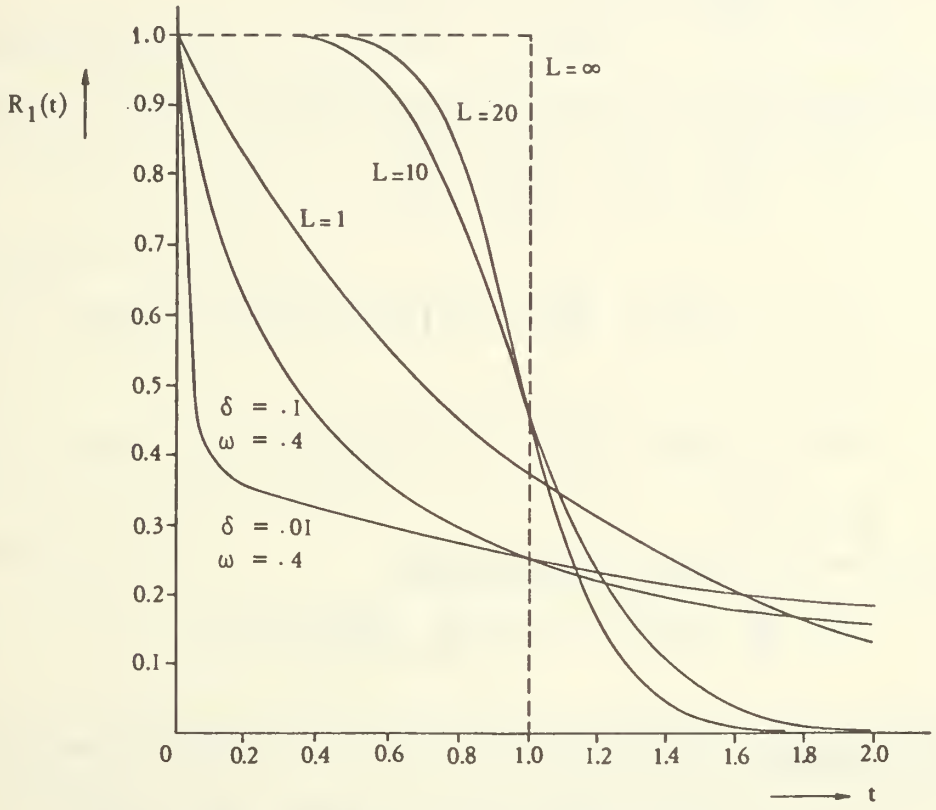


FIGURE 3. Erlang and hyperexponential distributions.

4. THE RELIABILITY $R_{k+1}(t)$ IN CASE OF THE HYPER-EXPONENTIAL DISTRIBUTION

The probability density function $a(t)$ of the hyper-exponential distribution is given by

$$a(t) = \delta\mu\omega e^{-\delta\mu t} + (1-\delta)\mu(1-\omega)e^{-(1-\delta)\mu t}.$$

If $\delta = 1/2$ or $\omega = 0$ or $\omega = 1$, $a(t)$ becomes the density function of an exponential distribution. Results for $R_{k+1}(t)$ being known in this case, we only study the other values of δ and ω.

We write for $a(t)$ the somewhat shorter notation

$$a(t) = ce^{-at} + de^{-bt} \text{ with } a = \delta\mu, \quad b = (1-\delta)\mu,$$
$$c = \delta\mu\omega, \quad d = (1-\delta)\mu(1-\omega).$$

Taking the Laplace transform of $a(t)$:

$$La(s) = \frac{c}{s+a} + \frac{d}{s+b}$$

so

$$Lf(s) = \sum_{j=0}^{k+1} \binom{k+1}{j} \left(\frac{c}{s+a}\right)^{k+1-j} \left(\frac{d}{s+b}\right)^{j}.$$

Making some preparations for the inverse transformation, yields:

$$Lf(s) = \sum_{\alpha=1}^{k+1} \frac{A_\alpha}{(s+a)^\alpha} + \sum_{\beta=1}^{k+1} \frac{B_\beta}{(s+b)^\beta},$$

with

$$A_\alpha = \frac{1}{(a-b)^{k+1-\alpha}} \sum_{l=1}^{k+1-\alpha} \binom{k+1}{l} \binom{k-\alpha}{l-1} (-1)^l c^{k+1-l} d^l, \quad \alpha = 1, \ldots, k,$$

$$A_{k+1} = c^{k+1},$$

$$B_\beta = \frac{1}{(b-a)^{k+1-\beta}} \sum_{l=1}^{k+1-\beta} \binom{k+1}{l} \binom{k-\beta}{l-1} (-1)^l d^{k+1-l} c^l, \quad \beta = 1, \ldots, k,$$

$$B_{k+1} = d^{k+1}.$$

Actually, taking the inverse Laplace transform, we find

$$f(t) = \sum_{\alpha=1}^{k+1} \frac{A_\alpha}{(\alpha-1)!} t^{\alpha-1} e^{-at} + \sum_{\beta=1}^{k+1} \frac{B_\beta}{(\beta-1)!} t^{\beta-1} e^{-bt}.$$

Integrating $f(t)$ we find the reliability $R_{k+1}(t)$ of a single unit with k-spares in case of a hyper-exponential failure density function per unit:

$$R_{k+1}(t) = \sum_{\alpha=1}^{k+1} \left[\frac{A_\alpha}{a^\alpha} e^{-at} \left\{ \sum_{h=0}^{\alpha-1} \frac{(at)^h}{h!} \right\} \right] + \sum_{\beta=1}^{k+1} \left[\frac{B_\beta}{b^\beta} e^{-bt} \left\{ \sum_{h=0}^{\beta-1} \frac{(bt)^h}{h!} \right\} \right] \tag{4.1}$$
$$= \sum_{j=0}^{k} \left\{ \sum_{i=j+1}^{k+1} \frac{A_i}{a^i} \right\} e^{-at} \frac{(at)^j}{j!} + \sum_{j=0}^{k} \left\{ \sum_{i=j+1}^{k+1} \frac{B_i}{b^i} \right\} e^{-bt} \frac{(bt)^j}{j!}.$$

Application and combination of the above mentioned results yields the following formula for $R_{k+1}(t)$:

$$R_{k+1}(t) = \sum_{j=0}^{k} c_j \frac{(at)^j}{j!} e^{-at} + \sum_{j=0}^{k} d_j \frac{(bt)^j}{j!} e^{-bt}, \tag{4.2}$$

with

$$c_j = \omega^{k+1}\left\{1 + \sum_{i=j+1}^{k}\left[\left(\frac{\delta}{2\delta - 1}\right)^{k+1-i}\sum_{l=1}^{k+1-i}\binom{k+1}{l}\binom{k-i}{l-1}(-1)^l\left(\frac{(1-\delta)(1-\omega)}{\delta\omega}\right)^l\right]\right\},$$

$$d_j = (1-\omega)^{k+1}\left\{1 + \sum_{i=j+1}^{k}\left[\left(\frac{1-\delta}{1-2\delta}\right)^{k+1-i}\sum_{l=1}^{k+1-i}\binom{k+1}{l}\binom{k-i}{l-1}(-1)^l\left(\frac{\delta\omega}{(1-\delta)(1-\omega)}\right)^l\right]\right\},$$

$$j = 0, 1, \ldots, k-1,$$

$$c_k = \omega^{k+1}, \quad d_k = (1-\omega)^{k+1}.$$

5. THE RELIABILITY $R_{k+1}(t)$ IN CASE OF THE ERLANG DISTRIBUTION

Differentiation of the Erlang reliability function with parameters L and λ yields the following expression for the probability density function of the lifetime of a single unit:

$$a(t) = \frac{(L\lambda)^L}{(L-1)!}\, t^{L-1}e^{-L\lambda t}.$$

Taking the Laplace transform gives

$$La(s) = \frac{(L\lambda)^L}{(s + L\lambda)^L},$$

hence

$$Lf(s) = \frac{(L\lambda)^{(k+1)L}}{(s + L\lambda)^{(k+1)L}}.$$

We must take the inverse transform of the expression for $f(t)$:

$$f(t) = \frac{(L\lambda)^{(k+1)L}}{[(k+1)L - 1]!}\, t^{(k+1)L-1}\, e^{-L\lambda t},$$

again a p.d.f. of an Erlang distribution with parameters $(k+1)L$ and $\frac{\lambda}{k+1}$. Thus, we find for the reliability $R_{k+1}(t)$ of a single unit which k spare parts in case of the Erlang distribution:

$$R_{k+1}(t) = e^{-L\lambda t}\sum_{i=0}^{(k+1)L-1}\frac{(L\lambda t)^i}{i!}. \tag{5.1}$$

6. A NUMERICAL EXAMPLE

To applicate the formula of Section 4 we study a system of units where all the units have a hyper-exponentially distributed lifetime. We assume for the parameters the following values: $\omega = 0.7$, $\delta = 0.01$, and $\mu = 50$. This means that each unit has a mean lifetime of 1.4 with a standard deviation of 1.9.

If we have to allocate spares to the system so that a 99% reliability exists at $t = 10$ we see from Figure 4 that 18 spares are sufficient. If we base computations on the exponential distribution with the same mean lifetime of 1.4 we find that 14 spares already give the same reliability at $t = 10$. And if 14 spares are allocated to this system where the units have an hyperexponential distribution only a 96% reliability at $t = 10$ can be given. Generally, the hyperexponential distribution is more conservative in estimating number of spares than the exponential distribution with the same mean.

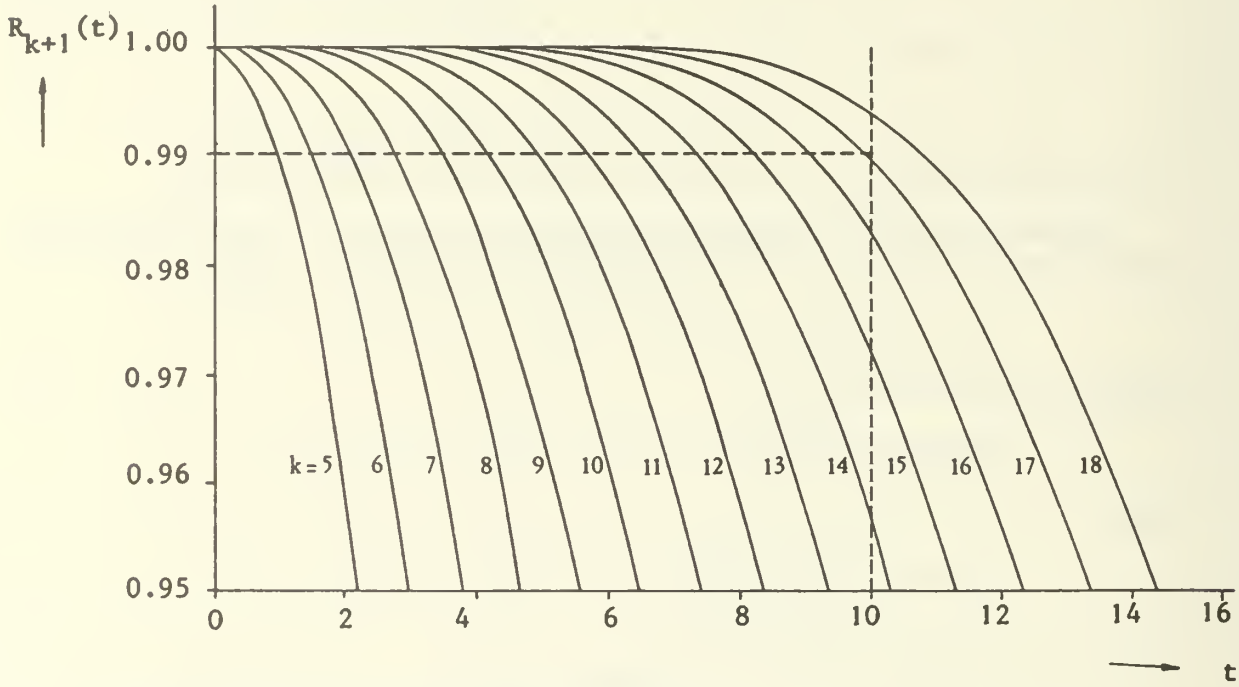


FIGURE 4. Graphs of $R_{k+1}(t)$ in case of hyperexponential distribution with $\omega = .7$ $\delta = .01$ and $\mu = 50$.

REFERENCES

[1] Barlow, R.E. and F. Proschan, *Mathematical Theory of Reliability* (John Wiley and Sons, New York, 1965).

[2] Cox, D.R., *Renewal Theory* (Methuen and Co. Ltd., London, 1962).

[3] Kapur, K.C. and L.R. Lamberson, *Reliability in Engineering Design* (John Wiley and Sons, New York, 1977).

[4] Marshall, A.W. and F. Proschan, "Classes of Distributions Applicable in Replacement with Renewal Theory Implications," Proceedings, Sixth Berkeley Symposium, 1970.

[5] Palm, C., Intensitätsschwankungen im Fernspechverkeher, Ericsson Technics 44, 1-89 (1943).

[6] Parzen, E., *Modern Probability Theory and Its Applications* (John Wiley and Sons, New York, 1970).

[7] Ruiz-Pala, E., Avila-Beloso, C., and W.W. Hines, *Waiting-Line Models* (Reinhold Publishing Corporation, New York, (1967).

[8] Shooman, M.L., *Probabilistic Reliability: An Engineering Approach* (McGraw-Hill Book Company, New York, 1968).
[9] Van Dam, J., "Optimization of Spare Part Kits," Doctoral Thesis, University of Technology, Delft, Netherlands (1975).

SELECTING AMONG WEIBULL, LOGNORMAL AND GAMMA DISTRIBUTIONS USING COMPLETE AND CENSORED SAMPLES

Siswadi and C. P. Quesenberry

North Carolina State University
Raleigh, North Carolina

ABSTRACT

In a recent paper, Kent and Quesenberry [19] considered using certain optimal invariant statistics to select the best fitting member of a collection of probability distributions using complete samples of life data. In the present work extensions of this approach in two directions are given. First, selection for complete samples based on scale and shape invariant statistics is considered. Next, the selection problem for type I censored samples is considered, and both scale invariant and maximum likelihood selection procedures are studied. The two-parameter (scale and shape) Weibull, lognormal, and gamma distributions are considered and applications to real data are given. Results from a (small) comparative simulation study are presented.

1. INTRODUCTION

The two-parameter (scale and shape) Weibull, lognormal, and gamma distributions are all commonly used in reliability and life testing problems. The problem of selecting one of these three distributions for a particular sample, either complete or censored, is a difficult one. In this work we consider basing the selection on the values of certain selection statistics computed from the sample. Although we consider selection based only on sample information, it should be noted that in some practical problems further information may be available which should also be weighed in the final selection of a distribution. Such information could be derived, for example, from known physical characteristics of a failure mechanism and its failure rate.

Also, it should be mentioned that throughout this paper we treat the selection of one of the distributions as a forced selection problem. That is, we formally select exactly one of the competing distributions. In practice, we may sometimes wish to use a partial selection procedure which does not necessarily always select one distribution over the others when they are close together, as indicated by selection statistics that are nearly equal. The selection statistics posed here may be readily used in partial selection schemes. However, we do not explicitly consider these applications in the present paper.

Kent and Quesenberry [19], KQ, proposed a forced selection procedure based on statistics that are invariant under scale transformations. Other relevant literature includes a paper by Dumonceaux, Antle, and Haas [11] who examined maximum likelihood ratio (MLR) tests for discriminating between two models with unknown location and scale parameters, and compared empirically the power of MLR tests with that of uniformly most powerful invariant (UMPI) tests for discriminating between normal and Cauchy distributions. They actually recommend

the MLR test over the UMPI test on the basis of relatively good power and ease of computation. Dumonceaux and Antle [10] gave an MLR procedure for discriminating between Weibull and lognormal distributions that is based on the fact that the logarithms of both Weibull and lognormal random variables have location-scale parameter distributions. In a recent paper, Bain and Engelhardt [2] considered a likelihood ratio selection statistic for selecting between gamma and Weibull distributions.

Some graphical procedures for the selection problem have been given by Nelson [21], and by Barlow and Campo [3]. Other papers that are related to the present work include Hogg, Uthoff, Randles, and Davenport [18], who discuss a number of selection procedures, including one based on location-scale invariant statistics; Dyer ([12],[13]) who considers a number of selection procedures for discriminating between pairs of classes of location-scale distributions; and Uthoff ([24],[25]) who considers some particular invariant statistics. As general references for invariant tests see Hajek and Sidak [15] and Lehmann [20], and for MLR tests see Cox [8]. Volodin [26] considers a generalized three-parameter gamma distribution and discriminates between two-parameter gamma and Weibull distributions by making scale invariant tests on the other parameters.

As mentioned above, KQ considered selecting among the gamma, lognormal and Weibull families for the complete sample problem. The selection statistic posed was formed by first deriving a scale invariant statistic that is optimal in the sense that it minimizes the sum of the two probabilities of selecting the incorrect distribution for two conformable distributions (cf. Quesenberry and Starbuck [22]); and then replacing the shape parameter by its ML estimator. Such procedures were called *suboptimal,* and the selection statistics for the three families were set out in simple closed form formulas in that paper.

In the present work we consider two major changes in the approach and problem considered in KQ. First, for the lognormal and Weibull families we use optimal *scale and shape* invariant selection statistics. Also, we consider the selection problem for type I censored samples as well as for complete samples. For these cases the selection statistics are generally expressed as definite integrals whose evaluation requires numerical integration. Thus, a substantial part of this work has necessarily been concerned with the development of computer algorithms to evaluate these integrals.

2. DENSITIES AND SELECTION FUNCTIONS

In many applied problems it is reasonable to assume that the location parameters of life distributions are known. Thus, we consider distributions with only scale and shape parameters unknown. The densities of the gamma, lognormal and Weibull distributions to be studied are given in Table 1. The symbol $I_{(a,b)}(x)$ in Table 1, and elsewhere, is the indicator function of the interval $(a,\ b)$, i.e., $I_{(a,b)}(x) = 1$ if $a < x < b$, and is zero, otherwise.

We consider type I censored samples, which are obtained when a number of items are put on life test and observed for a previously specified fixed time T. Thus, the parent density for the observed lives is a truncated version of the complete samples density given in Table 1.

The approach used here is of the same general form as that in KQ. A *selection statistic, S,* is defined for each of the three parametric classes, and the class with the largest selection statistic is chosen as the best fitting family for a given sample.

TABLE 1 — *Densities of Weibull, Lognormal, and Gamma Distributions*

Name	Symbol	Density	
Weibull	$W(\theta, \beta)$	$f_1 = \frac{\beta}{\theta}\left(\frac{x}{\theta}\right)^{\beta-1} \exp[-(x/\theta)^{\beta}] \cdot I_{(0,\infty)}(x);$	$\theta, \beta > 0$
Lognormal	$LN(\theta, \sigma)$	$f_2 = \frac{1}{\sqrt{2\pi}\,\sigma x} \cdot \exp\{-[\ln(x/\theta)]^2/2\sigma^2\} \cdot I_{(0,\infty)}(x);$	$\theta, \sigma > 0$
Gamma	$G(\theta, \alpha)$	$f_3 = \theta^{-\alpha}[\Gamma(\alpha)]^{-1}x^{\alpha-1}\exp(-x/\theta) \cdot I_{(0,\infty)}(x);$	$\theta, \alpha > 0$

We consider some transformation properties of these distributions before defining selection statistics for them. If X is a random variable with either a $G(\theta, \alpha)$, $LN(\theta, \sigma)$ or $W(\theta, \beta)$ distribution, then consider the transformation

$$Y = aX^b, \quad a > 0, \quad b > 0. \tag{2.1}$$

If X is a $W(\theta, \beta)$ random variable, then Y is a $W(a\theta^b, 3/b)$ random variable; and if X is a $LN(\theta, \sigma)$ random variable, then Y is a $LN(a\theta^b, b\sigma)$ random variable. Thus, Weibull random variables are transformed to Weibull random variables by (2.1), and lognormal random variables are transformed to lognormal random variables by this transformation. Unfortunately for present purposes at least, the gamma distributions do not share this property since a gamma random variable is not always transformed to another gamma random variable by (2.1). That is to say, the $G(\theta, \alpha)$ class is not a scale-shape class that is *conformable* with the lognormal and Weibull classes as defined by Quesenberry and Starbuck [22]. Nevertheless, we use here a selection statistic for the complete samples problem which is essentially the value of the density function of a maximal invariant when each of the three parents is assumed.

For $x_1, \ldots, x_n$ an observed sample, we define the selection statistic for a density function f_i $(i = 1, 2, 3)$ by

$$S_i = \int_0^{\infty}\int_0^{\infty} f_i(\gamma x_1^{\lambda}, \ldots, \gamma x_n^{\lambda})\gamma^{n-1}\lambda^{n-2}(x_1 \ldots x_n)^{\lambda} d\gamma\, d\lambda. \tag{2.2}$$

Due to the property of the $G(\theta, \alpha)$ distribution discussed above, the selection statistic S_3 of (2.2) is a function of the parameter α. We obtain a selection statistic by replacing α by its maximum likelihood estimator, $\hat{\alpha}$, in this function. The selection statistics used in this work are given in Table 2.

TABLE 2 — *Scale-shape Invariant Selection Statistics for Complete Samples*

Family	i	S_i
$W(\theta, \beta)$	1	$\Gamma(n)\int_0^{\infty} (\prod x_i)^{\lambda}(\Sigma x_i^{\lambda})^{-n}\lambda^{n-2}d\lambda$
$LN(\theta, \sigma)$	2	$1/2\; n^{-1/2}\pi^{-1/2(n-1)}\Gamma[1/2(n-1)][\Sigma \ln^2 x_i - (\Sigma \ln x_i)^2/n]^{1/2(n-1)}$
$G(\theta, \hat{\alpha})$	3	$\Gamma(n\hat{\alpha})\Gamma^{-n}(\hat{\alpha})\int_0^{\infty} (\prod x_i)^{\hat{\alpha}\lambda}(\Sigma x_i^{\lambda})^{-n\hat{\alpha}}\lambda^{n-2}d\lambda$

The evaluation of these functions requires numerical integration, except for the lognormal selection function. It is often computationally easier to compute and compare the logarithms of the selection statistics than the statistics themselves. To estimate the parameter α of the gamma distribution, we use the ML estimator $\hat{\alpha}$ of Greenwood and Durand [14] which has been studied further by Bowman and Shenton [6], and was recently used and given in detail by KQ. The selection procedures proposed here are closely related to uniformly most powerful invariant (UMPI) tests for separate families testing problems. For the particular case of selecting between lognormal and Weibull distributions, using the selection statistics of Table 2 is equivalent to using the UMPI test statistic for classifying a sample into one of these two distributions. If α_1 denotes the probability that a sample from a lognormal parent will be classified as a Weibull sample, and α_2 the probability that a Weibull sample will be classified as a lognormal sample, then the above selection procedure will minimize $\alpha_1 + \alpha_2$ among all procedures invariant with respect to the transformations of (2.1). Or, if the probability of selecting each distribution is 1/2, then this procedure minimizes the total probability of misclassification, *viz.*, $(\alpha_1 + \alpha_2)/2$ (see KQ, Section 3).

3. SIMULATION RESULTS FOR COMPLETE SAMPLES

In this section we report results of a simulation study of the selection rules proposed above.

In order to obtain results that can be compared with those of KQ, samples were generated from nine parent distributions: $W(1, 1/2)$, $W(1, 2)$, $W(1, 4)$, $LN(1, 0.4)$, $LN(1, 1)$, $LN(1, 2.5)$, $G(1, 1/2)$, $G(1, 2)$, and $G(1, 5)$ for $n = 10, 20, 30$. One thousand samples were generated from each of these distributions. The pairwise selection error rates are given in Table 3 and the observed rates of correct classification in the 3-way procedure are given in Table 5.

The entries in Table 3 are read as follows, using the first set of results as an example. The selection is to be made between W and LN families. One thousand samples of size 10 were generated from a W distribution, and 28 percent of these samples were classified as being from an LN distribution. One thousand samples of size 10 were generated from an LN distribution and 36 percent of these samples were classified as being from a W distribution. Note that since W and LN are conformable scale-shape families, these results do not depend upon which particular members of the families that are involved.

This procedure has total error probabilities for the case of a lognormal *vs* a Weibull that are the smallest possible for a scale-shape invariant procedure. Comparison of the results for this case given in Table 3 and those in Table 4 of KQ shows that the present optimal procedure has very little advantage over the suboptimal procedure of KQ. For the other two-way selection problems, gamma *vs* lognormal and gamma *vs* Weibull, the comparisons of the procedures of this paper with those of KQ are not clearcut. This is because the gamma distribution does not admit a UMPI statistic with respect to the transformations of (2.1).

In view of these observations we recommend the selection procedures set out in KQ on the grounds that (i) the selection statistics in Table 2 of KQ have convenient formulas that are readily evaluated, and (ii) the error rates achieved by those procedures appear to be about as favorable as for those achieved by the much more computationally difficult scale-shape invariant procedures.

TABLE 3 — *Selection Error Rates for Pairwise Procedures**

	n	$X \sim W$		$X \sim LN$	Total	
	10	.36		.28	.32	
	20	.21		.23	.22	
	30	.15		.18	.16	
n	$X \sim G(1/2)$	$X \sim W(1/2)$	Total	$X \sim G(1/2)$	$X \sim W(2)$	Total
10	.40	.42	.41	.40	.44	.42
20	.37	.38	.38	.37	.36	.37
30	.34	.35	.34	.34	.33	.34
	$X \sim G(1/2)$	$X \sim W(4)$		$X \sim G(2)$	$X \sim W(1/2)$	
10	.40	.40	.40	.44	.42	.43
20	.37	.29	.33	.42	.38	.40
30	.34	.23	.29	.40	.35	.37
	$X \sim G(2)$	$X \sim W(2)$		$X \sim G(2)$	$X \sim W(4)$	
10	.44	.44	.44	.44	.40	.42
20	.42	.36	.39	.42	.29	.36
30	.40	.33	.36	.40	.23	.31
	$X \sim G(5)$	$X \sim W(1/2)$		$X \sim G(5)$	$X \sim W(2)$	
10	.39	.42	.41	.39	.44	.41
20	.31	.38	.35	.31	.36	.34
30	.30	.35	.32	.30	.33	.31
	$X \sim G(5)$	$X \sim W(4)$				
10	.39	.40	.39			
20	.31	.29	.30			
30	.30	.23	.26			
n	$X \sim G(1/2)$	$X \sim LN(.04)$		$X \sim G(1/2)$	$X \sim LN(1)$	
10	.25	.41	.33	.25	.33	.29
20	.15	.38	.27	.15	.27	.21
30	.09	.35	.22	.09	.21	.15
	$X \sim G(1/2)$	$X \sim LN(2.5)$		$X \sim G(2)$	$X \sim LN(0.4)$	
10	.25	.23	.24	.37	.41	.39
20	.15	.14	.15	.31	.38	.34
30	.09	.08	.08	.24	.35	.29
	$X \sim G(2)$	$X \sim LN(1)$		$X \sim G(2)$	$X \sim LN(2.5)$	
10	.37	.33	.35	.37	.23	.30
20	.31	.27	.29	.31	.14	.22
30	.24	.21	.22	.24	.08	.16
	$X \sim G(5)$	$X \sim LN(0.4)$		$X \sim G(5)$	$X \sim LN(1)$	
10	.44	.41	.42	.44	.33	.38
20	.36	.38	.37	.36	.27	.32
30	.33	.35	.34	.33	.21	.27
	$X \sim G(5)$	$X \sim LN(2.5)$				
10	.44	.23	.33			
20	.36	.14	.25			
30	.33	.08	.20			

*See Section 3 for explanation of table entries and discussion of these results.

TABLE 4 — *Total Error Rates for Gamma vs Weibull for Likelihood Ratio Procedure of Bain and Engelhardt*

α	n	β:	.5	2	4
.5	10		.435	.405	.385
	20		.375	.345	.320
2	10		.470	.440	.420
	20		.415	.385	.360

Also, Bain and Engelhardt [2] give in their Table 2 some probabilities of correct selection between gamma and Weibull distributions using a likelihood ratio test statistic. Their results can be used to construct total error rates comparable to those of Table 3, for a few selected values of the gamma and Weibull shape parameters. We have computed these values and give them in Table 4. Comparison of the total error rates of Tables 3 and 4 shows no trend in favor of either procedure.

Table 5 gives the selection rates in our simulation study for the three-way scale-shape invariant selection procedure. The entries in in Table 5 are read as follows, using the first set of results as an example. The selection is to be made among the gamma, Weibull and lognormal distributions. One thousand samples of size 10 were generated from a $G(1, 1/2)$ distribution, of which 57 percent are classified as G, 21 percent are classified as W and 23 percent are classified LN. The results in Table 5 can be compared with the results in Table 5 of KQ. The comparisons do not show that either of these procedures has a clear advantage, however, the selection procedure of KQ may have a slight edge. Thus, as for the two-way selection procedures above, we favor the computationally simpler scale invariant, SI, procedure of KQ.

TABLE 5 — *Selection Rates for Three-Way Procedure*

n	$X \sim G(1/2)$			$X \sim G(2)$			$X \sim G(5)$		
	G	W	LN	G	W	LN	G	W	LN
10	.57	.21	.23	.20	.44	.37	.18	.39	.44
20	.62	.26	.12	.28	.43	.31	.32	.31	.36
30	.66	.29	.05	.37	.40	.24	.38	.30	.33
	$X \sim W(1/2)$			$X \sim W(2)$			$X \sim W(4)$		
	G	W	LN	G	W	LN	G	W	LN
10	.42	.23	.35	.15	.57	.29	.16	.60	.24
20	.38	.41	.21	.24	.64	.12	.19	.71	.10
30	.35	.51	.15	.26	.67	.07	.18	.77	.05
	$X \sim LN(0.4)$			$X \sim LN(1)$			$X \sim LN(2.5)$		
	G	W	LN	G	W	LN	G	W	LN
10	.17	.24	.59	.14	.19	.67	.14	.15	.72
20	.23	.15	.62	.18	.09	.73	.04	.19	.77
30	.26	.10	.65	.16	.05	.79	.02	.16	.82

4. SELECTION WITH CENSORING

Suppose that from a random sample of size n on a parent random variable with density and distribution functions f and F, respectively, only the values less than a prespecified time T are observed. If r is the number of values less than T, then r is a binomial rv with probability function $b(r; F(T), n)$. Let $x_1, \ldots, x_r$ be the observed values, indexed in the same order as the original sample, and $x_{(1)}, \ldots, x_{(r)}$ be the corresponding order statistics. We require selection procedures based on the values $x_1, \ldots, x_r$, and r. We have studied procedures based on scale-shape invariance, as considered above for complete samples, scale invariance as in KQ, and maximum likelihood ratio procedures. Of these procedures, only the SI and maximum likelihood, ML, procedures will be described now, since these procedures will be recommended for reasons given below.

When f and F are functions of parameters θ, say f_θ and F_θ, the likelihood function corresponding to $x_1, \ldots, x_r$ and r is

$$L_\theta(x_{(1)}, \ldots, x_{(r)}, r; T) =$$

(4.1)

$$\{n!/(n-r)!\}\{1 - F_\theta(T)\}^{n-r} \prod_{i=1}^{r} f_\theta(x_{(i)}) I_{(0,T)}(x_{(i)}).$$

The scale invariant selection statistic is defined by

(4.2) $$S = \int_0^\infty L(\lambda x_{(1)}, \ldots, \lambda x_{(r)}, r; \lambda T)\lambda^{r-1} d\lambda,$$

where the scale parameter in L has been set equal to one, and S depends upon a shape parameter. The shape parameter in S for each of the three families considered here will be replaced by its maximum likelihood estimator, obtained by maximizing the likelihood in (4.1).

We also consider selection for the censored case using, essentially, a likelihood ratio procedure. In this approach we use the maximum value of the likelihood function in (4.1) as the selection statistic. Formally, the selection statistic is

$$S = \sup_\theta \{L_\theta(x_{(1)}, \ldots, x_{(r)}, r; T)\},$$

(4.3)

$$= L_{\hat{\theta}}(x_{(1)}, \ldots, x_{(r)}, r; T),$$

for $\hat{\theta}$ the ML estimator(s) of θ.

The selection functions for these two methods for the three families are given in Table 6.

We have written programs to evaluate the selection statistics of Table 6. A brief description of this work follows in the remainder of this section. For more detail see Siswadi [23].

Maximum likelihood estimates for the scale and shape parameters are required for the ML selection functions, and for the shape parameter for the SI selection functions. For the Weibull class, these estimates were obtained as solutions of the ML equations in Cohen [7].

TABLE 6 — *Selection Statistics for Censored Samples*

Family	i	S_i
Scale Invariant, *SI*		
$W(\theta, \beta)$	1	$\frac{n!}{(n-r)!}\Gamma(r)\hat{\beta}^{r-1}\left(\prod_{i=1}^{r} x_{(i)}\right)^{\hat{\beta}-1}\left(\sum_{i=1}^{r} x_{(i)}^{\hat{\beta}} + (n-r)T^{\hat{\beta}}\right)^{-r}$
$LN(\theta, \sigma)$	2	$\frac{n!}{(n-r)!}(\sqrt{2\pi}\,\hat{\sigma})^{1-r}r^{-1/2}\left(\prod_{i=1}^{r} x_{(i)}\right)^{-1}\exp\left\{-\frac{1}{2\hat{\sigma}^2}\cdot\left[\sum_{i=1}^{r}(\ln x_{(i)}^2) - \frac{1}{r}\left(\sum_{i=1}^{r}\ln x_{(i)}\right)^2\right]\right\}\cdot E\left[1-\phi\left(\frac{u+\ln T}{\hat{\sigma}}\right)\right]^{n-r}$ where u is a $N\left(-\sum_{i=1}^{r}\ln x_{(i)}/r,\ \hat{\sigma}^2/r\right)$ r.v.
$G(\theta, \alpha)$	3	$\frac{n!}{(n-r)!}\Gamma(r\hat{\alpha})\Gamma^{-r}(\hat{\alpha})\left(\prod_{i=1}^{r} x_{(i)}\right)^{\hat{\alpha}-1}\left(\sum_{i=1}^{r} x_{(i)}\right)^{-r\hat{\alpha}} E\left[1-\Gamma\left(Tu \Big/ \sum_{i=1}^{r} x_{(i)},\ \hat{\alpha}\right)/\Gamma(\hat{\alpha})\right]^{n-r}$ where u is a $G(1, r\hat{\alpha})$ r.v.
Maximum Likelihood, *ML*		
$W(\theta, \beta)$	1	$\frac{n!}{(n-r)!}\hat{\beta}^r\hat{\theta}^{-r\hat{\beta}}\left(\prod_{i=1}^{r} x_{(i)}\right)^{\hat{\beta}-1}\exp\left[-\hat{\theta}^{\hat{\beta}}\left(\sum_{i=1}^{r} x_{(i)}^{\hat{\beta}} + (n-r)T^{\hat{\beta}}\right)\right]$
$LN(\theta, \sigma)$	2	$\frac{n!}{(n-r)!}\left[1-\phi\left(\frac{\ln T - \ln\hat{\theta}}{\hat{\sigma}}\right)\right]^{n-r}\left(\prod_{i=1}^{r} x_{(i)}\right)^{-1}(\sigma\sqrt{2\pi})^{-r}\exp\left\{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{r}(\ln x_{(i)} - \ln\hat{\theta})^2\right\}$
$G(\theta, \alpha)$	3	$\frac{n!}{(n-r)!}\Gamma^{-n}(\hat{\alpha})\left[\Gamma(\hat{\alpha}) - \Gamma\left(\frac{T}{\hat{\theta}},\ \hat{\alpha}\right)\right]^{n-r}(\hat{\theta})^{-r\hat{\alpha}}\left(\prod_{i=1}^{r} x_{(i)}\right)^{\hat{\alpha}-1}\exp\left(-\frac{1}{\hat{\theta}}\sum_{i=1}^{r} x_{(i)}\right)$

Note: $\Gamma(a, b) = \int_0^a s^{b-1}\exp(-s)ds$, ϕ is $N(0,1)$ *df*.

For the lognormal class solutions for the ML equations were obtained using results of Harter and Moore [17], adjusted for type I censoring. Another procedure for lognormal type I censored samples is given by Aitchison and Brown [1]. Solutions of the ML equations for the gamma class were obtained from the results of Harter and Moore [16], adjusted for type I censoring. After the ML estimates of the scale and shape parameters are obtained, the evaluations of the ML selection functions of Table 6 are straightforward.

After the ML estimate $\hat{\beta}$ of the shape parameter of the Weibull distribution is obtained, the SI selection function is readily evaluated. However, the selection functions for both the lognormal and gamma scale invariant procedures are difficult to evaluate, and we have used Monte Carlo and importance sampling from the normal distribution and gamma distribution, respectively (see Davis and Rabinowitz [9]) to evaluate them.

5. SIMULATION RESULTS FOR CENSORED SAMPLES

We have conducted a small Monte Carlo simulation study of the two selection methods discussed above for censored samples to provide some information on the error rates for these procedures. These empirical error rates allow comparisons of the two procedures with each other as well as with complete sample rates given in Section 3 and in KQ. Comparison with complete sample rates gives a measure of the loss of information due to censoring.

The families of distributions considered were $W(1/2)$, $W(4)$, $G(1/2)$, $G(2)$, $LN(0.4)$, and $LN(2.5)$; and the sample size was $n = 30$ in all cases. The truncation point T was chosen so that the df, $F(T) = 0.90$, i.e., for a mean rate of 10% censoring. One hundred samples were generated for each of the above distributions except $W(4)$, for which 16 samples were generated. The running time for some cases was very long and this limited the number of samples that could be generated.

The misclassification rates for pairwise selection are given in Table 7, which is comparable to Table 3 for the complete samples case. Note in Table 7 that the ML and SI procedures give similar error rates for *W vs LN* and, in fact, both give the same total error rate of 0.29. Comparison of these results with those of the SI procedure for complete samples (see Table 4 of KQ) shows that there is a rather large loss of information due to censoring since the *W*, *LN* and total error rates are 0.19, 0.15, and 0.17, respectively.

TABLE 7 — *Misclassification Rates for Pairwise Selection Procedures-Censored Sample (n = 30)*

		Procedure	$X \sim W$	$X \sim LN$	Total	
		ML	.40	.18	.29	
		SI	.45	.12	.29	
Procedure	$X \sim G(1/2)$	$X \sim W(1/2)$	Total	$X \sim G(1/2)$	$X \sim W(4)$	Total
ML	.38	.32	.35	.38	.44	.41
SI	.52	.16	.34	.52	.50	.50
	$X \sim G(2)$	$X \sim W(1/2)$		$X \sim G(2)$	$X \sim W(4)$	
ML	.36	.32	.34	.36	.44	.40
SL	.39	.16	.28	.39	.50	.45
	$X \sim G(1/2)$	$X \sim LN(0.4)$		$X \sim G(1/2)$	$X \sim LN(2.5)$	
ML	.14	.32	.23	.14	.08	.11
SI	.20	.31	.26	.20	.09	.15
	$X \sim G(2)$	$X \sim LN(0.4)$		$X \sim G(2)$	$X \sim LN(2.5)$	
ML	.35	.32	.34	.35	.08	.22
SI	.42	.31	.37	.42	.09	.21

For the two-way selection error rates in Table 7 that involve a gamma distribution, neither the ML nor the SI procedure appears to have an overall advantage. Also, by comparing these cases with the same cases in Table 4 of KQ, we feel that the loss of information due to censoring is not so large as for the *W vs LN* case commented on above.

The classification rates for three-way selection procedures are given in Table 8. Again, neither the ML nor the SI procedure appear to have any overall advantage, and both perform quite well. Also, by comparison with Table 5 of KQ it appears that ten percent censoring has little effect on the probability of correctly classifying a lognormal sample, but the probabilities of correctly classifying either Weibull or gamma samples are reduced somewhat.

TABLE 8 — *Classification Rates for Three-Way Procedures—Censored Sample (n = 30)*

Procedure	$X \sim G(1/2)$			$X \sim G(2)$			$X \sim W(1/2)$		
	G	*W*	*LN*	*G*	*W*	*LN*	*G*	*W*	*LN*
ML	.60	.29	.11	.29	.36	.35	.27	.35	.38
SI	.47	.35	.18	.28	.32	.40	.16	.40	.44
	$X \sim W(4)$			$X \sim LN(0.4)$			$X \sim LN(2.5)$		
ML	.13	.56	.31	.23	.09	.68	.01	.17	.82
SI	.25	.50	.25	.23	.08	.69	.01	.13	.86

6. A USER PROGRAM AND EXAMPLES FOR CENSORED SAMPLES

The selection procedures for the three families of distributions have been proprammed in FORTRAN. A listing of this program can be obtained from the authors. The program computes the selection statistics for complete and censored samples according to the formulas given in Tables 2 and 6, respectively.

For the scale invariant procedure, the selection statistics are computed by the Monte Carlo method given in Davis and Rabinowitz [9]. The program was tested on several examples and on many samples produced through simulation. In general, the selection statistics estimated did not appear reliable for heavily censored samples. Therefore, in the user program for the scale invariant procedure, the selection results are not printed if the coefficient of variation of the replicated values in the Monte Carlo method is larger than 35%.

EXAMPLE 1: Birnbaum and Saunders [5] considered a set of data of lifetimes, in thousands of cycles, of aluminum sheeting under periodic loading, to illustrate the gamma family. If we assume that the experiment was terminated at a prespecified time, say $T = 1900$, then the censored observations and the results of the selection procedure are presented in Table 9. For these data, the Weibull family is selected by both the ML and SI procedures, however, the selection statistics for the gamma family are only slightly smaller. It is also to be noted, although the details are not given here, for the complete sample the selection procedure based on the selection statistics given in Table 2 yields the same results.

TABLE 9 — *Results of Selection Procedure*

Lifetimes of Aluminum Sheeting under Periodic Loading					
370	706	716	746	785	797
844	855	858	886	886	930
960	988	990	1000	1010	1016
1018	1020	1055	1085	1102	1102
1108	1115	1120	1134	1140	1199
1200	1200	1203	1222	1235	1238
1252	1258	1262	1269	1270	1290
1293	1300	1310	1313	1315	1330
1355	1390	1416	1419	1420	1420
1450	1452	1475	1478	1481	1485
1502	1505	1513	1522	1522	1530
1540	1560	1567	1578	1594	1602
1604	1608	1630	1642	1674	1730
1750	1750	1763	1768	1781	1782
1792	1820	1868	1881	1890	1893
1895					
Sample Size = 101 Sample Observed = 91 Truncation point = 1900					

Maximum Likelihood Estimates		
Family	Scale	Shape
Weibull	0.154149D+04	0.404114D+01
Gamma	0.125214D+03	0.112550D+02
Lognormal	0.135159D+04	0.317034D+00

Family	Selection Statistic
Maximum Likelihood Procedure	
Weibull	0.147164D+03
Gamma	0.146470D+03
Lognormal	0.144092D+03
The Family Selected is Weibull	
Scale Invariant Procedure	
Weibull	0.144432D+03
Gamma	0.144103D+03
Lognormal	0.141762D+03
The Family Selected is Weibull	

EXAMPLE 2: Bartholomew [4, p. 370] gave the failure times of 15 items that failed during a specified period of testing from an original sample of size $n = 20$. He states that the items have an exponential life distribution, and uses the exponential distribution to perform analyses of the data. We have used these data in the selection program, and the results are given in Table 10. Both the maximum likelihood and scale invariant procedures prefer the lognormal distributions, which casts some doubt on the assumption of an exponential parent distribution.

TABLE 10 — *Results of Selection Procedure*

Bartholomew Data					
3	19	23	26	27	37
38	41	45	58	84	90
99	109	138			
Sample size = 20					
Sample Observed = 15					
Truncation point = 150					

Maximum Likelihood Estimates		
Family	Scale	Shape
Weibull	0.105498D 03	0.108289D 01
Gamma	0.876146D 02	0.116892D 01
Lognormal	0.682517D 02	0.122585D 01

Family	Selection Statistic
Maximum Likelihood Procedures	
Weibull	−0.669185D 01
Gamma	−0.664251D 01
Lognormal	−0.660273D 01
The Family Selected is Lognormal	
Scale Invariant Procedure	
Weibull	−0.720101D 01
Gamma	−0.699896D 01
Lognormal	−0.667347D 01
The Family Selected is Lognormal	

REFERENCES

[1] Aitchison, J. and J.A.C. Brown, *The Lognormal Distribution* (Cambridge University Press, London, 1957).

[2] Bain, L.J. and M. Engelhardt, "Probability of Correct Selection of Weibull vs Gamma Based on Likelihood Ratio," Communications in Statistics, *A9,* 375-381 (1980).

[3] Barlow, R.E. and R. Campo, "Total Time on Test Processes and Applications to Failure Data Analysis," in *Reliability and Fault Tree Analysis: Theoretical and Applied Aspects of System Reliability and Safety Assessment,* 451-481 (SIAM, Philadelphia, 1975).

[4] Bartholomew, D.J., "The Sampling Distribution of an Estimate Arising in Life Testing," Technometrics, *5,* 361-374 (1963).
[5] Birnbaum, Z.W. and S.C. Saunders, "A Statistical Model for Life-length of Materials," Journal of the American Statistical Association, **53,** 151-160 (1958).
[6] Bowman, K.O. and L.R. Shenton, *Properties of Estimator for the Gamma Distribution,* Report CTC-1, Union Carbide Corporation, Nuclear Division, Oak Ridge, TN (1968).
[7] Cohen, A.C., "Maximum Likelihood Estimation in the Weibull Distribution Based on Complete and on Censored Samples," Technometrics, *7,* 579-588 (1965).
[8] Cox, D.R., "Tests of Separate Families of Hypotheses," in *Proceedings of the Fourth Berkeley Symposium 1,* 105-123 (University of California Press, Berkeley, 1961).
[9] Davis, P.J. and P. Rabinowitz, *Methods of Numerical Integration* (Academic Press, New York, 1975).
[10] Dumonceaux, R. and C.E. Antle, "Discrimination between the Lognormal and the Weibull Distributions," Technometrics, *15,* 923-926 (1973).
[11] Dumonceaux, R., C.E. Antle, and G. Haas, "Likelihood Ratio Test for Discrimination between Two Models with Unknown Location and Scale Parameters," Technometrics, *15,* 19-27 (1973).
[12] Dyer, A.R., "Discrimination Procedures for Separate Families of Hypotheses," Journal of the American Statistical Association, *68,* 970-974 (1973).
[13] Dyer, A.R., "Hypotheses Testing Procedures for Separate Families of Hypotheses," Journal of the American Statistical Association, *69,* 140-145 (1974).
[14] Greenwood, J.A. and D. Durand, "Aids for Fitting the Gamma Distribution by Maximum Likelihood," Technometrics, *2,* 55-65 (1960).
[15] Hajek, J. and Z. Sidak, *Theory of Rank Tests* (Academic Press, New York, 1967).
[16] Harter, H.L. and A.H. Moore, "Maximum Likelihood Estimation of the Parameters of Gamma and Weibull Populations from Complete and from Censored Samples," Technometrics, *7,* 639-643 (1965).
[17] Harter, H.L. and A.H. Moore, "Local-Maximum-Likelihood Estimation of the Parameters of Three-Parameter Lognormal Populations from Complete and Censored Samples," Journal of the American Statistical Association, *61,* 841-842 (1966).
[18] Hogg, R.V., V.A. Uthoff, R.H. Randles, and A.S. Davenport, "On the Selection of the Underlying Distribution and Adaptive Estimation," Journal of the American Statistical Association, *67,* 597-600 (1972).
[19] Kent, J. and C.P. Quesenberry, "Selecting Among Probability Distributions Used in Reliability," Technometrics, *24,* 59-65 (1982).
[20] Lehmann, E.L., *Testing Statistical Hypotheses* (John Wiley and Sons, New York, 1959).
[21] Nelson, W., "Theory and Applications of Hazard Plotting for Censored Failure Data," Technometrics, *14,* 945-966 (1972).
[22] Quesenberry, C.P. and R.R. Starbuck, "On Optimal Tests for Separate Hypotheses and Conditional Probability Integral Transformations," Communications in Statistics: Theory and Methods, Part A, *5,* 507-524 (1976).
[23] Siswadi, "Selecting Among Weibull, Lognormal and Gamma Distributions for Complete and Censored Samples," Unpublished North Carolina State University Thesis (1981).
[24] Uthoff, V.A., "An Optimum Test Property of Two Well-Known Statistics," Journal of the American Statistical Association, *65,* 1597-1600 (1970).
[25] Uthoff, V.A., "The Most Powerful Scale and Location Invariant Test of the Normal vs the Double Exponential," Annals of Statistics, *1,* 170-174 (1973).
[26] Volodin, I.N., "On the Discrimination of Gamma and Weibull Distributions," Theory of Probability and its Applications, *19,* 383-393 (1974).

A TIME SERIES ANALYSIS OF SOME INTERRELATED LOGISTICS PERFORMANCE VARIABLES*

Nozer D. Singpurwalla

The George Washington University
Washington, D.C.

Carlos M. Talbott**

Air Force Institute of Technology
Wright Patterson AFB, Ohio

ABSTRACT

This paper is a case study. We show how the powerful methods of time series analysis can be used to investigate the interrelationships between Alert Availability, a logistics performance variable, and Flying Hours, an operational requirement, in the presence of a major change in operating procedures and using contaminated data. The system considered is the fleet of C-141 aircraft of the U.S. Air Force. The major change in operating procedures was brought about by what is known as Reliability Centered Maintenance, and the contaminated data were due to anomalies in reporting procedures. The technique used is a combination of transfer function modeling and intervention analysis.

1. INTRODUCTION AND SUMMARY

In January 1976, the U.S. Air Force began some experimental modifications to the existing maintenance policies for the fleet of C-141 aircraft. These modifications were a part of a Department of Defense project known as "Reliability Centered Maintenance," henceforth denoted by RCM. The modifications involved an *extension* of the maintenance intervals and a *reduction* in the amount of scheduled maintenance. The experimental phase of the project ended in June 1977, and the modified policies were officially and permanently instituted at that time. The anticipated benefit from RCM was a decrease in scheduled maintenance activity, with a consequent increase in "alert availability." *Alert availability,* henceforth denoted by AA, is the instantaneous probability that a typical aircraft is available to react to an execution order. In practice, its average value, over say a month, is computed by dividing the monthly total number of "fleet operational hours" by the total number of "fleet available hours."† A plot of AA from October 1973 through November 1979 is shown in Figure 1.

*This research was initiated and supported by Air Force Logistics Management Center, Contract F01600-79-D0146, Office of Naval Research Contract N00014-77-C-0263 and U.S. Army Research Office Grant DAAG-29-80-C-0067.
**Major, USAF
† *Fleet available hours* is the cumulative time of possession across a fleet of aircraft, whereas *fleet operational hours* is the cumulative amount of operational time (i.e., up-time) for the fleet.

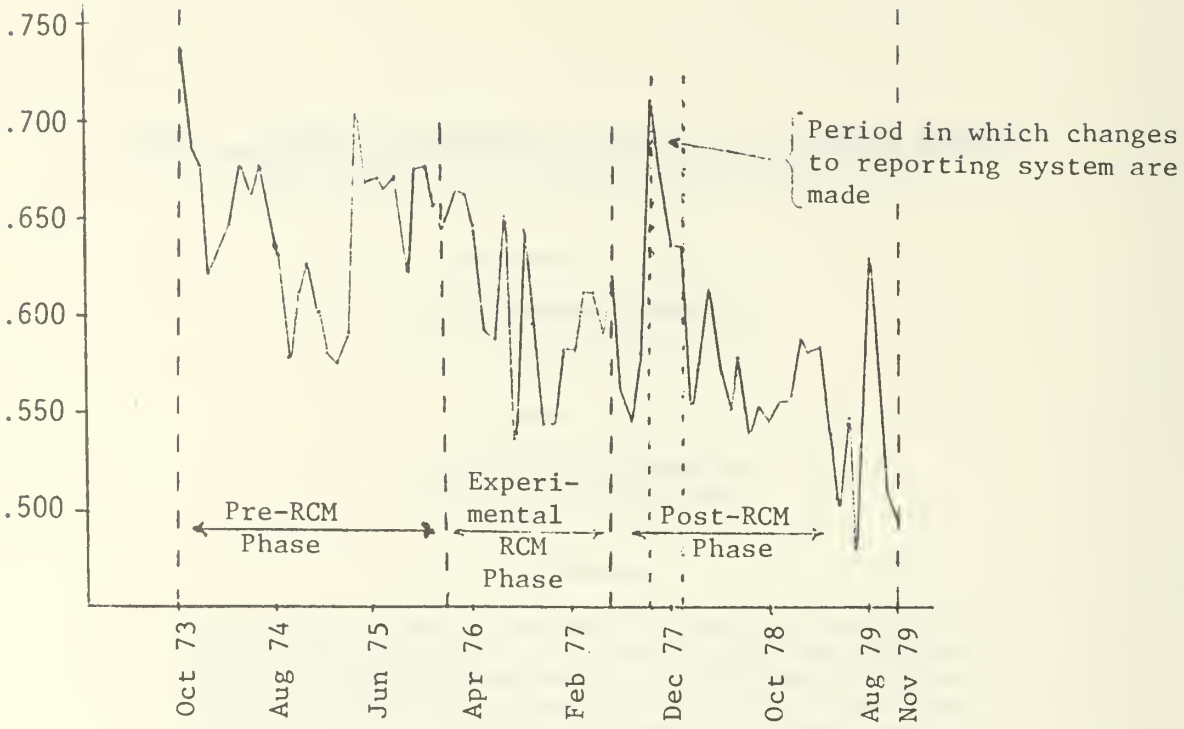


FIGURE 1. A plot of C-141 alert availablity from October 1973 to November 1979

In a previous study (Singpurwalla and Talbott [5]), we investigated the effects of RCM on several variables which describe what is known as the "logistics performance" of the fleet. Our conclusion was that there was no evidence of an improvement in the logistics performance of the fleet due to RCM; on the contrary, in some cases there was a clear indication of deterioration in performance. These conclusions were particularly true of AA, which is considered to be an important logistics performance variable. One criticism of this previous study, and a valid one, is that it did not take into consideration the influence of other "operational variables," which in addition to RCM may affect the logistics performance variables. An important operational variable, which is suspected of being strongly related to the AA, is "flying hours." *Flying hours* is the total number of hours flown by the fleet of the C-141's over a certain period of time, say one month. A plot of the monthly flying hours, from October 1973 through November 1979, is shown in Figure 2. Another criticism of our previous study, and again a valid one, is that it did not adequately account for the fact that some of the AA data were "messy." Specifically, there were some revisions to the information system for reporting the operational status of the aircraft that resulted in some possible anomalies in the reported values of the AA.

The analysis that is described here was initiated with a view towards rectifying the limitations of our previous study. Here, using the AA and the flying hours as examples, we demonstrate how a procedure for investigating the interrelationships between the two, using messy data and RCM as an intervention, can be developed and used. The approach we take is a combination of those described in Box and Jenkins [1], pp. 335-420) for the analysis of multiple time series, and in Box and Tiao [2] for intervention analysis. Our conclusion is that the AA is

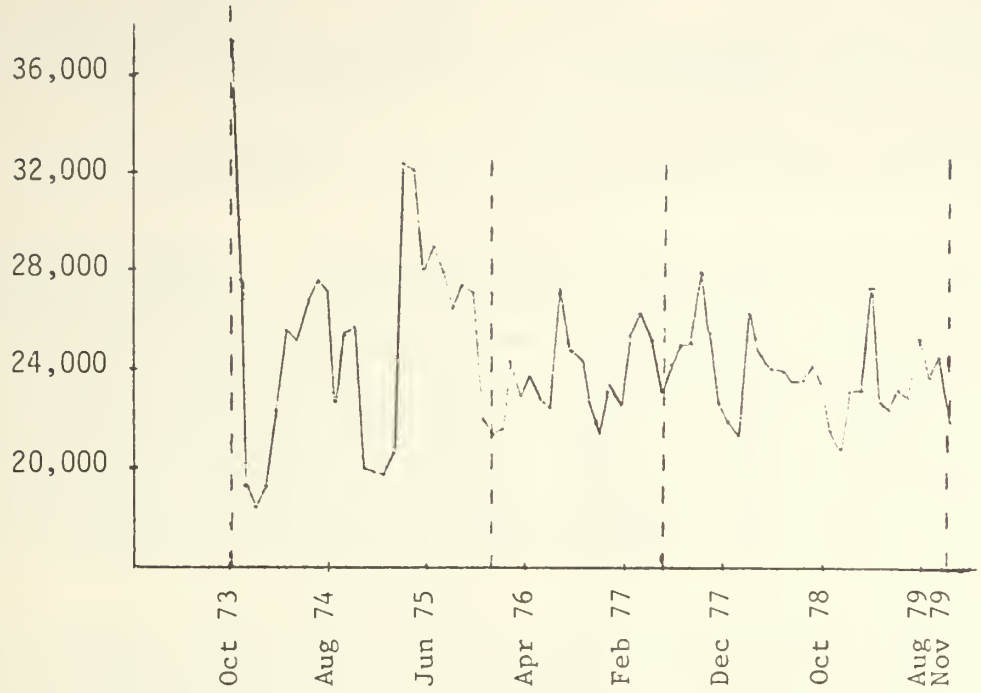


FIGURE 2. A plot of C-141 flying hours from October 1973 to November 1979

indeed related to the flying hours, as has been conjectured, but that even after taking this relationship into consideration, and in the presence of messy data, our previous conclusion still holds, namely, that there is no clear evidence of improvement in the AA after the initiation of RCM.

Before going into the details of our analysis, we want to emphasize that the above conclusion and its practical implications are not intended to be the main theme of this paper. Rather, our aim is to suggest and to demonstrate how the powerful methods of time series analysis can be used to analyze the messy and interrelated data that often arise in a study of the reliability and the logistics performance of large military systems.

In what follows, we presume that the reader has a knowledge of autoregressive integrated moving average (ARIMA) processes and is familiar with the notation, terminology, and methodology described in Box and Jenkins [1].

2. NOTATION, PRELIMINARIES, AND AN OUTLINE OF THE PROCEDURE

Let X and Y be two variables of interest, and let X_t and Y_t be their values at time t. In our case, we let X_t denote the total flying hours for the fleet of C-141's during the tth month, and Y_t the alert availability during that month; t varies from October 1973 through November 1979. A sequence of values $X_t, X_{t-1}, \ldots$ will be denoted by $\{X_t\}$.

To discern the relationship between X_t and Y_t, we strive to obtain a *linear transfer function model* (Box and Jenkins [1], p. 379) of the form

$$(1 - \delta_1 B - \ldots - \delta_r B^r) Y_t = (w_0 - w_1 B - \ldots - w_s B^s) X_{t-b} + N_t, \tag{2.1}$$

where $B^m X_t = X_{t-m}$, $m = 0, 1, \ldots,$ and $\delta_1, \ldots, \delta_r, w_0, \ldots, w_s$ are unknown constants to be estimated; b represents the lag of Y_t with respect to X_t; and N_t denotes a noise component described by a suitable ARIMA process.

Equation (2.1) describes the alert availability at time t in terms of the previous values of the alert availability and the present and previous values of the flying hours.

Following the procedures suggested by Box and Jenkins [1] and by Haugh and Box [3], we first consider the sequence (series) $\{X_t\}$ and transform it to a "white noise series," $\{a_t\}$, by way of an appropriate univariate time series model. Similarly, we also reduce the sequence $\{Y_t\}$ to a white noise series, $\{b_t\}$, using an appropriate "intervention analysis model" (see Box and Tiao [2]). Such a model is necessary here in order to account for RCM as well as for the presence of some anomalous observations in $\{Y_t\}$. This procedure of reducing $\{X_t\}$ and $\{Y_t\}$ to $\{a_t\}$ and $\{b_t\}$, respectively, is known as *prewhitening* the respective series. We then cross correlate the two prewhitened sequences $\{a_t\}$ and $\{b_t\}$ in order to obtain an indication of the relationship between $\{X_t\}$ and $\{Y_t\}$. The cross correlation function is used to suggest values for $r,s,$ and b in (2.1). Finally, we use the prewhitened sequences $\{a_t\}$ and $\{b_t\}$ to estimate the constants $\delta_1, \ldots, \delta_r, w_0, \ldots, w_s$ for some selected values for r, s, and b.

Regarding the prewhitening of $\{Y_t\}$ using a model for intervention analysis, we remark that intervention due to RCM can be described by a sequence of indicator variables, say $\{Z_t\}$, where Z_t takes a value of 0 for all t representing the months prior to January 1976, and a value 1 for all t thereafter. Recall that January 1976 is the date of intervention—the date at which experimental RCM was initiated. The response of the sequence $\{Y_t\}$, the *output sequence* denoting AA, to the *input sequence* $\{Z_t\}$ can take various functional forms. These are depicted in Figure 3; they have been taken from Box and Tiao [2]. We strive to use the most appropriate form of the response function for the situations at hand.

3. PREWHITENING THE FLYING HOURS SERIES $\{X_t\}$

An inspection of Figure 2 finds that the fluctuations of the series about its mean changes over time. In such situations it is common to take the natural logarithms of the data. Accordingly, our analysis of flying hours involves a logarithmic transformation of the original data.

We find that flying hours can best be described by an ARIMA $(2,0,0) * (2,0,0)_4$ process. However, the residuals from this model reveal significant autocorrelations at lags 5, 10, 15, ..., etc. Consequently, we fit an ARIMA $(0,0,0) * (0,0,1)_5$ model to these residuals. Based upon these considerations, an appropriate model for prewhitening can be written as

$$(3.1) \qquad (1 - .54B + .11B^2)(1 + .38B^4 + .37B^8)(X_t - 10.09) = (1 + .18B^5)a_t.$$

This model produces residuals whose autocorrelation function (ACF) and log spectral density are shown in Figure 4, in (a) and (b), respectively. These residuals are listed in Table 1.

4. PREWHITENING THE ALERT AVAILABILITY SERIES $\{Y_t\}$

In our previous study we considered alert availability as two separate series, one ending prior to January 1976 and the other beginning with January 1976. We found that the loga-

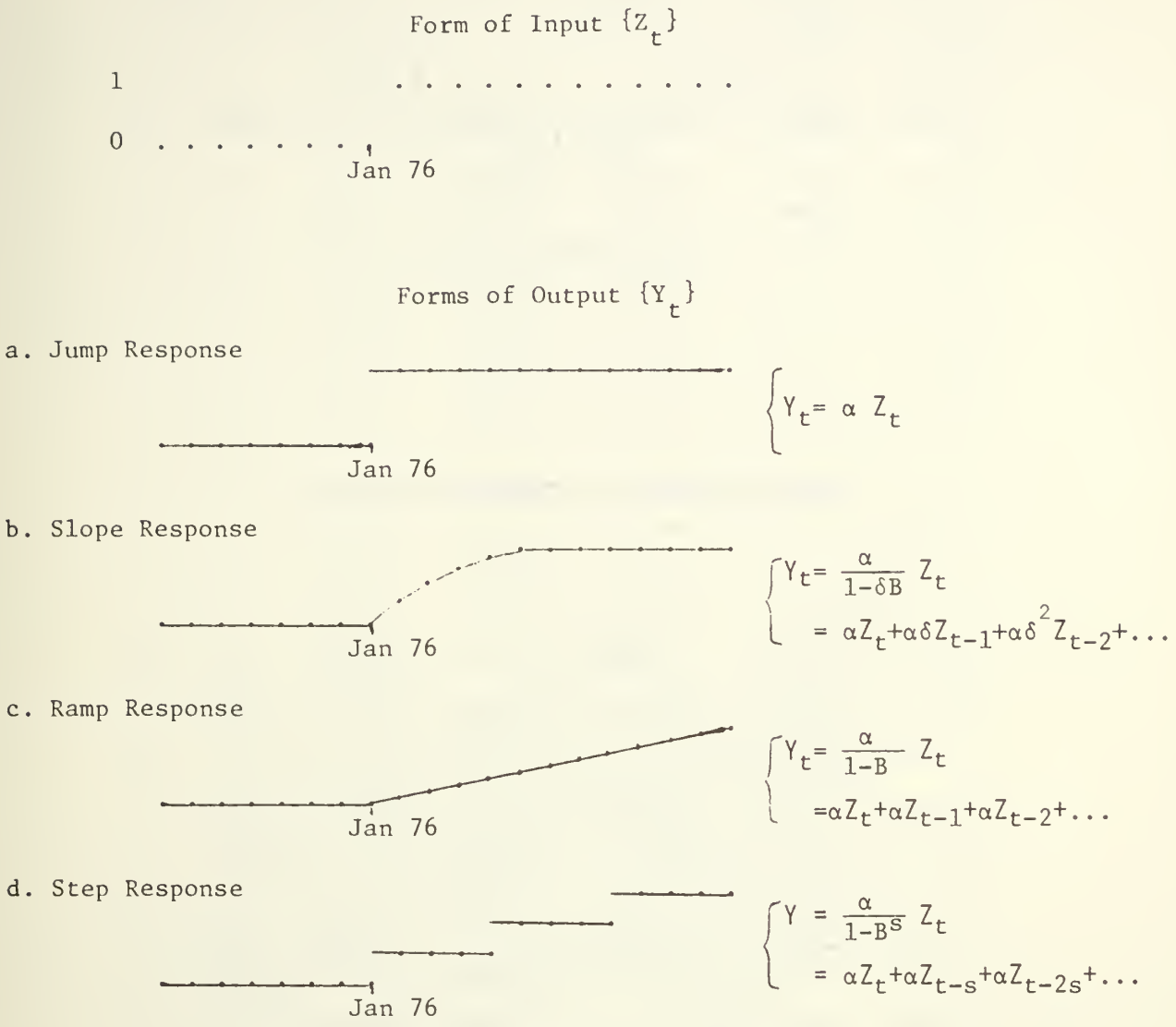
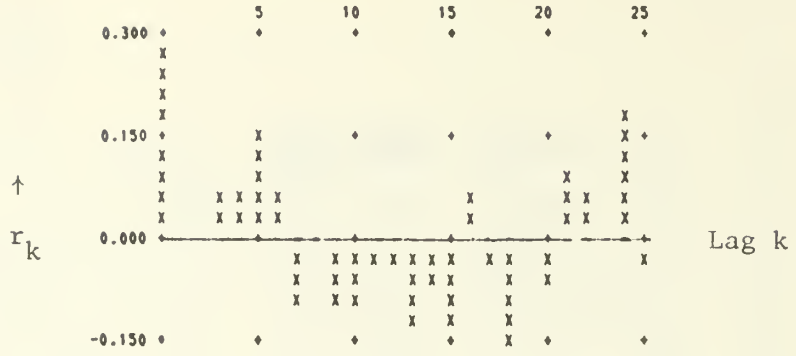


FIGURE 3. Typical forms of the response $\{Y_t\}$ to the input $\{Z_t\}$



(a) Graph of observed series ACF

(b) Graph of log spectral density

FIGURE 4. Behavior of the ACF and the log of the spectral density of the residuals from an ARIMA $(0,0,0) \cdot 0,0,1)_5$ model fit to residuals from an ARIMA $(2,0,0) \cdot 2,0,0)_4$ model

TABLE 1 — *Residuals from Prewhitening Models for Flying Hours and Alert Availability*

TIME T	A (T)	B (T)	TIME T	A (T)	B (T)
JAN 1975	−.131303		JUL 1977	.037152	−.053020
FEB 1975	−.026086	−.074950	AUG 1977	.022717	.012580
MAR 1975	−.034914	−.003150	SEP 1977	.021623	−.045700
APR 1975	.283323	.108000	OCT 1977	.080163	.053860
MAY 1975	.060148	.026820	NOV 1977	.037052	.039220
JUN 1975	−.032649	.056160	DEC 1977	−.048706	−.063430
JUL 1975	.135207	.085220	JAN 1978	−.047455	−.011080
AUG 1975	.124353	.105500	FEB 1978	−.058318	.016660
SEP 1975	−.007629	.039270	MAR 1978	.130172	−.010630
OCT 1975	.059647	−.003746	APR 1978	−.053118	.013570
NOV 1975	.046232	.035070	MAY 1978	−.008318	.001221
DEC 1975	.020893	.007938	JUN 1978	.020092	−.023900
JAN 1976	−.025669	.033680	JUL 1978	.036141	.067210
FEB 1976	−.021861	.054060	AUG 1978	−.073053	−.013030
MAR 1976	.109077	.095080	SEP 1978	−.002367	−.015090
APR 1976	−.101662	−.033110	OCT 1978	−.050561	−.099890
MAY 1976	.001382	−.081130	NOV 1978	−.054985	.001532
JUN 1976	−.044392	−.057940	DEC 1978	−.110742	.036330
JUL 1976	−.012095	.068650	JAN 1979	.041226	.051630
AUG 1976	.068336	−.175700	FEB 1979	−.043154	−.043154
SEP 1976	−.047820	.075940	MAR 1979	.106237	.047260
OCT 1976	−.035721	−.047270	APR 1979	−.173297	−.027790
NOV 1976	−.066278	−.066570	MAY 1979	.007886	−.090300
DEC 1976	−.044366	−.042450	JUN 1979	−.031905	.030150
JAN 1977	.003072	−.040260	JUL 1979	−.014306	−.183300
FEB 1977	−.073529	.076220	AUG 1979	−.027671	.163600
MAR 1977	.045873	−.014810	SEP 1979	−.030462	.029880
APR 1977	.097139	.062520	OCT 1979	.026290	−.062220
MAY 1977	.008354	.098570	NOV 1979	−.070995	−.015390
JUN 1977	−.089749	.137600			

VOL. 29, NO. 4, DECEMBER 1982

rithms of the observations in both these series could best be described by an ARIMA $(1,0,0) * (1,1,0)_6$ model. Here, we prewhiten the *entire* series by considering the onset of RCM as an intervention and incorporating this effect into the ARIMA $(1,0,0) * (1,1,0)_6$ model.

Figure 1 indicates clearly that the AA had been decreasing after January 1976, the onset of RCM. However, the rate of decrease of the AA during the experimental phase of RCM is greater than the rate of decrease during the post-RCM phase. This suggests that the effect of RCM during the experimental phase may be different from the effect after the experimental phase. One possible reason for this difference is that there was a piecemeal introduction of RCM, air base by air base, during the experimental phase, along with several trial revisions in the maintenance policies. We also notice some large spikes in value of the AA during October, November, and December 1977 and January 1978. A cause for these large values may be that the information system for reporting the operational status of aircraft was revised during this period. The AA may have been artificially increased during this period due to anomalies of reporting. Thus, it appears that we need three distinct components to our intervention analysis model. We represent these components by three indicator variables, J_t, K_t, and L_t, where:

$$J_t = \begin{cases} 1 & \text{for Jan 76} \leqslant t \leqslant \text{May 77} \\ 0 & \text{otherwise} \end{cases}$$

$$K_t = \begin{cases} 1 & \text{for Jun 77} \leqslant t \leqslant \text{Sep 77} \\ 1 & \text{for Feb 78} \leqslant t \leqslant \text{Nov 79} \\ 0 & \text{otherwise} \end{cases}$$

$$L_t = \begin{cases} 1 & \text{for Oct 77} \leqslant t \leqslant \text{Jan 78} \\ 0 & \text{otherwise} \end{cases}$$

Note that the union of the three sequences $\{J_t\}$, $\{K_t\}$, and $\{L_t\}$ constitutes the series $\{Z_t\}$, defined earlier.

As for the functional form of the response (series $\{Y_t\}$) to these components (see Figure 3), we remark that ramp and step responses appear to be possible candidate forms for $\{J_t\}$ and $\{K_t\}$, whereas a jump response is appropriate for $\{L_t\}$. We exclude a jump response for the first two because it would not account for the gradual decline evident in the data, and we consider a slope response to be inappropriate due to the absence of a sustained leveling of the data. We conjecture that a step response might have a period of three months, which coincides with the length of time between scheduled minor inspections. Our aim now is to compare the functional forms of the ramp response versus the step response in order to determine which combination might best represent the relationship between AA and the series $\{Z_t\}$.

We fit all reasonable combinations of response forms with the variables J_t, K_t, and L_t, and find that step type responses, along with the original ARIMA model, i.e., a model of the form

$$Y_t = \frac{\alpha_1}{1 - B^3} J_t + \frac{\alpha_2}{1 - B^3} K_t + \alpha_3 L_t + N_t, \tag{4.1}$$

where N_t denotes an ARIMA $(1,0,0) * (1,1,0)_6$ process, produces residuals that are most satisfactory (see Figure 5). The model (4.1) describes the series $\{Y_t\}$ as a step function with three-month incremental decreases of size α_1 during the experimental RCM phase, as a step function with three-month incremental decreases of size α_2 after implementation of RCM, and as a jump function with height α_3 during the October 1977-January 1978 time frame. Thus we have the prewhitening model

$$(4.2) \qquad (1 - .18B)(1 + .35B^6)(1 - B^6)\,Y_t = \frac{-.02}{1 - B^3} J_t + \frac{-.01}{1 - B^3} K_t + .15L_t + b_t.$$

We remark that the estimated parameters α_1 and α_2 are negative, as expected, with α_1 greater than α_2. Also, the estimated parameter α_3 is positive, which is intuitively apparent. We also remark that in the residual series (see Figure 5) there remain some significant auto-correlations at lag 12. We can account for this by fitting an ARIMA $(0,0,0)*(2,0,0)_{12}$ model to the residuals. However, this would provide us with only 34 observations with which to perform a cross correlation analysis. Since paucity of data can lead us to questionable results,* we elect to continue working with the residuals of the model (4.2). These residuals are given in Table 1 as values of b_t; also given there are values of a_t, the prewhitened flying hours.

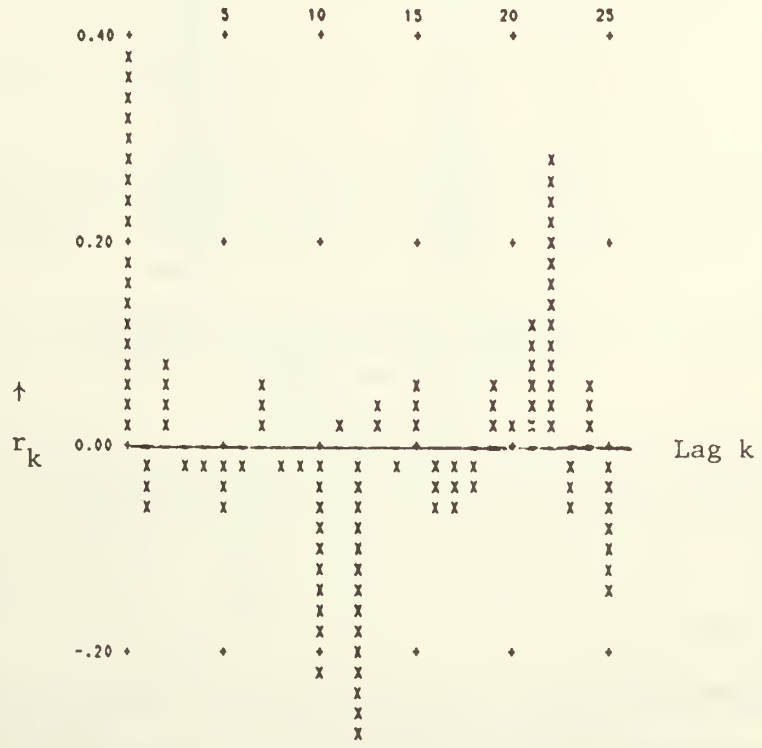


FIGURE 5. Intervention model for alert availability; graph of the ACF of the residual series

*Actually, we did perform a cross correlation analysis with only 34 observations and obtained results that were counter-intuitive.

5. CROSS CORRELATION ANALYSIS

Using the prewhitened series $\{a_t\}$ and $\{b_t\}$, we now develop the transfer function model (2.1). We remark that the $\{a_t\}$ series with its 59 observations must be modified by omitting the January 1975 observation so that it will match the $\{b_t\}$ series, which has 58 observations.

Following procedures of Box and Jenkins and using the estimated "impulse response weights," V_k, shown in Figure 6, we remark that possible values of (r,s,b) in (2.1) are (1,3,0) or (2,3,0). From the autocorrelation functions of the generated noise series (see Singpurwalla and Talbott [4]), we remark that a possible model for the noise N_t is an ARIMA (1,0,1).

K	V(K)
0	0.224
1	0.260
2	0.035
3	0.126
4	0.026
5	0.000
6	0.078
7	0.127
8	-0.025
9	0.023
10	-0.016
11	0.149
12	-0.227
13	-0.213
14	-0.021
15	0.146

FIGURE 6. Estimated impulse response weights $V(K)$ from a cross correlation of prewhitened flying hours and alert availability

We fit both the (2,3,0) and the (1,3,0) transfer function models with an ARIMA (1,0,1) noise component to the data and find that the more parsimonious (1,3,0) model results in better residuals. This estimated model, whose residuals have an ACF as shown in Figure 7, is

$$(5.1) \qquad (1 + .53B)b_t = (.25 + .51B + .24B^2 + .1B^3)a_t + \frac{1 - .11B}{1 + .12B}e_t,$$

where e_t is random noise.

6. THE COMBINED INTERVENTION ANALYSIS TRANSFER FUNCTION MODEL

We can expand the transfer function model (5.1) for the prewhitened series $\{a_t\}$ and $\{b_t\}$ by using the following relationships from (3.1) and (4.2):

$$(6.1) \qquad a_t = \frac{(1 - .54B + .11B^2)(1 + .38B^4 + .37B^8)(\tilde{X}_t)}{1 + .18B^5}$$

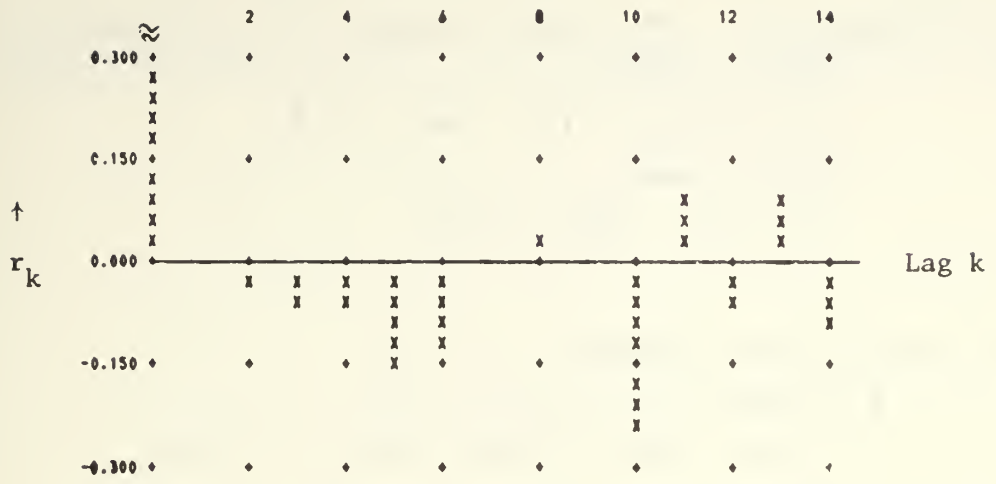


FIGURE 7. A (1,3,0) transfer function model between flying hours and alert availability; graph of the ACF of the residual series

$$b_t = (1 - .18B)(1 + .35B^6)(1 - B^6)Y_t + \frac{.02}{1 - B^3}J_t + \frac{.01}{1 - B^3}K_t - .15L_t, \tag{6.2}$$

where $\tilde{X}_t = (X_t - 10.09)$.

Substituting (6.1) and (6.2) for a_t and b_t in Equation (5.1), we obtain a multiplicative transfer function model relating the input series $\{X_t\}$ to the output series $\{Y_t\}$ as

$$\begin{aligned}(1 + .18B^5)(1 + .53B)(1 - .18B)(1 + .35B^6)(1 - B^6)(1 + .12B)Y_t \\ = (.25 + .51B + .24B^2 + .1B^3)(1 - .54B + .11B^2)(1 + .38B^4 + .37B^8) \\ \cdot (1 + .12B)\tilde{X}_t - \frac{(1 + .12B)(1 + .53B)}{(1 + .18B^5)(1 - B^3)}[.02J_t] \\ - \frac{(1 + .12B)(1 + .53B)}{(1 + .18B^5)(1 - B^3)}[.01K_t] \\ + \frac{(1 + .12B)(1 + .53B)}{(1 + .18B^5)}[.15L_t] + (1 - .11B)e_t.\end{aligned} \tag{6.3}$$

Note that X_t and Y_t in (6.3) are in terms of the logarithms of the original data.

Expanding the polynomials and rewriting Equation (6.3) in conventional notation (in the interest of parsimony, coefficients with a value less than .1 are arbitrarily deleted), we can reduce the transfer function model to

$$
\begin{aligned}
Y_t = &-.47Y_{t-1} + .1Y_{t-2} - .18Y_{t-5} + .59Y_{t-6} + .31Y_{t-7} + .12Y_{t-11} \\
&+ .39Y_{t-12} + .17Y_{t-13} \\
&+ .25\tilde{X}_t + .68\tilde{X}_{t-1} + .11\tilde{X}_{t-3} + .26\tilde{X}_{t-5} + .25\tilde{X}_{t-9} \\
&+ .42e_{t-1} + .18e_{t-5} \\
&+ Z_t + .65Z_{t-1} - .18Z_{t-5} - .12Z_{t-6}
\end{aligned} \tag{6.4}
$$

where, following the notation of Section 4,

$$
\begin{aligned}
Z_t = \{&[-.02J_t - .02J_{t-3} - .02J_{t-6} - .02J_{t-9} - \ldots] \\
&+ [-.01K_t - .01K_{t-3} - .01K_{t-6} - .01K_{t-9} - \ldots] + .15L_t\}.
\end{aligned}
$$

7. CONCLUDING REMARKS

From Equation (6.4), we remark that the nonzero coefficients associated with the $\tilde{X}_t$'s imply that flying hours *do* have an effect on the alert availability in a manner specified by the functional form of the equation. Furthermore, an examination of Figure 2, together with the prewhitening transformation, Equation (3.1), reveals that there was no upward or downward trend in the flying hours during the period of study. However, there does appear to be a reduction in the variability of the flying hours as of the inception of RCM. In any case, it appears that Equation (6.4) supports the adage that "the more you fly, the less you fail," within limits.

The negative coefficients (albeit small) associated with the variables J_t and K_t in Equation (6.4) do, in the absence of any upward or downward trend in flying hours, support our premise that RCM has a tendency to reduce the alert availability.

The model, Equation (6.4), can be used not only for interpretative purposes as is done above, but within limits and with proper care, it can also be used to predict future availability given its previous values, and the present and previous values of the flying hours.

REFERENCES

[1] Box, G.E.P. and G.M. Jenkins, *Time Series Analysis: Forecasting and Control* (Holden-Day, San Francisco, 1976).

[2] Box, G.E.P. and G.C. Tiao, "Intervention Analysis with Applications in Economic and Environmental Problems," *Journal of the American Statistical Association*, 70-79 (March, 1975).

[3] Haugh, L.D. and G.E.P. Box, "Identification of Dynamic Regression (Distributed Lag) Models Connecting Two Time Series," *Journal of the American Statistical Association*, 121-130 (March, 1977)

[4] Singpurwalla, N.D. and C.M. Talbott, "A Combined Intervention Analysis/Transfer Function Model for Describing the Relationship Between C-141 Flying Hours and Alert Availability," AFLMC Report No. 790702-2, Air Force Logistics Management Center, Gunter AFB, Alabama (July, 1979)

[5] Singpurwalla, N.D. and C.M. Talbott, "An Assessment of the Benefits of Reliability Centered Maintenance on the C-141," Air Force Journal of Logistics, 11-15 (Winter, 1980)

VOL. 29, NO. 4, DECEMBER 1982

NONLINEAR INTEGER PROGRAMMING FOR VARIOUS FORMS OF CONSTRAINTS

Mary W. Cooper and Keyvan Farhangian

Department of Operations Research
and Engineering Management
School of Engineering and
Applied Science
Southern Methodist University
Dallas, Texas

ABSTRACT

A theoretical and computational investigation is made of the performance of a dynamic-programming-based algorithm for nonlinear integer problems with various types of constraints. We include linear constraints, aggregated linear constraints, separable nonlinear constraints and constraints involving maxima and minima. Separability of the objective function is assumed. The new feature of the algorithm is that two types of fathoming or pruning are used to reduce the size of tables and number of computations: fathoming by bounds and fathoming by infeasibility.

1. INTRODUCTION

Several lines of research have been suggested for mathematical programming problems which are nonlinear and also restricted to integer values for the decision variables. We define a very general notation for such a problem:

$$\text{(1)} \qquad \max \qquad z = \sum_{j=1}^{n} \phi_j(x_j)$$

subject to (P1)

$$\text{(2)} \qquad h_i(\bar{x}) \leqslant 0 \qquad i = 1m \ldots, m$$

$$\text{(3)} \qquad x_j \geqslant 0, \text{ integer } J = 1, \ldots, n$$

where $\phi_j(x_j)$ are nondecreasing functions and the constraint set is bounded and nonempty, and $\bar{x} = (x_1, x_2, \ldots, x_n)$.

Applications that have been or could be modeled by a representation of the form $(P1)$ include multidimensional knapsack problems [5], optimal allocation of scarce resources in production planning [5], servicing, repair and replacement of machinery [5], [7], [2], control of inventory [18], chemical reaction in a stirred tank [1], bottleneck problems in ecomomics [3], [1], [5], communication network problems [21], [4], models of growth and competition [1], [6], many aspects of chemical engineering [8], defensive missile allocation [33], and nonlinear

resource allocation [37], [38]. In the special case of binary variable problems there are applications to capital budgeting [29], sewage treatment and irrigation systems [28], media selection [36], cluster analysis [31], scheduling [19], [30], [2], location problems [35], [12], [32], and hydrological studies [12].

Cutting plane approaches to this problem have been suggested by Witzgall [34], and Granot, Granot, and Kallberg [16], among others. Solution methods based on enumeration or branch and bound have been given by Laughunn [25], Hansen [20], McBride and Yormark [26], and Korte, Krelle, and Oberhofer [22]-[24]. Other authors have taken the approach of linearizing to a linear integer programming problem and using existing integer algorithms. Papers by Glover [13], and Glover and Woolsey [14], [15], and Granot and Granot [17] lie in this class. The present paper is based on a fourth approach to nonlinear integer programming: using dynamic programming methodology while mitigating the traditional difficulty dynamic programming techniques have with multiply-constrained problems. This aim is referred to in the literature as reduction of state dimensionality [3].

2. HYPERSURFACE SEARCH

The hypersurface search method uses discrete, deterministic dynamic programming methodology to identify integer points on the hypersurface given by the objective function when the objective function value is set equal to some upper bound (in a maximization problem). Initially we "bypass" the original constraints of the problem and utilize dynamic programming recursion formulas for the following problem (P2) which is based only on the objective function of (P1):

$$\text{(4)} \qquad \max \quad z = \sum_{j=1}^{n} \phi_j(x_j)$$

subject to

$$\text{(5)} \qquad \sum_{j=1}^{n} \phi_j(x_j) \leqslant z_0 \qquad \text{(P2)}$$

$$\text{(6)} \qquad 0 \leqslant x_j \leqslant u_j \qquad j = 1, \ldots, n$$

$$\text{(7)} \qquad x_j, \text{ integer} \qquad j = 1, \ldots, n.$$

where z_0 is an upper bound on z and u_j are variable upper bounds. We complete the "forward pass" of dynamic programming. Then integer points are ordered according to decreasing value of the objective function using some efficient sort routine (we used Quicksort), and they are tested separately for infeasibility with the original constraints (2) of the problem (P1). Each candidate integer point is developed component by component in the usual "backward pass" of dynamic programming. But if we have one or more separable constraints we may detect infeasibility before completing the backward pass at the partial solution stage. Assume that in problem (P1), the inequalities (2) are replaced by

$$\text{(8)} \qquad \sum_{j=1}^{n} h_{ij}(x_j) \leqslant 0, \qquad i=1, \ldots, m.$$

Then a point can be eliminated as soon as any new component x_k^*, gives a partial sum

$$\text{(9)} \qquad \sum_{j=k}^{n} h_{ij}(x_j^*) \geqslant 0$$

for any constraint i. Here we are assuming a separable form for some or all of the constraints and, in addition, the nonnegativity of constraint coefficients. This algorithm is described in Cooper [9].

In a portfolio application [10] we have implemented several fathoming techniques used by Morin and Marsten [27] and others, and generalized by Denardo and Fox [11]. In order to use fathoming we must change the usual calculation of the recursive equations. It is replaced by a label-setting technique which has been called "reaching" in shortest path applications. This allows us to prune the entries in the dynamic programming tables. We eliminate partial solutions which would correspond to solutions with inferior values of the objective finction, and also partial solutions for which any complettion would be infeasible. The corresponding two types of fathoming are called fathoming by bounds and fathoming by infeasibility. The constraints must be separable, and be nondecreasing functions in order to get full computational effectiveness.

In this paper we extend the idea of fathoming by infeasibility from the portfolio application to the general problem with separable objective function and many different forms of constraints. We investigate linear, aggregated linear, separable, and nonseparable constraints.

3. LINEAR CONSTRAINTS

Computational testing was done on random problems with linear and squared terms in the objective function and linear constraints. The motivation for these tests was to find out whether fathoming by infeasibility with every constraint is superior to aggregating some of the constraints, and generating integer points, and then testing the solutions with the disaggregated constraints, as in [9]. Here are the two options tested:

(a) Fathoming by infeasibility with each constraint. Then all solutions at the end of the forward pass will be feasible with the original constraints. Sort according to objective function value to find the optimum.

(b) Fathoming by infeasibility with aggregated linear constraints. Several constraints are aggregated and the resulting constraint is used to eliminate some infeasible solutions in the forward pass. Sorting on objective function value is done and a backward pass through the dynamic programming tables tests solutions as their components are found for feasibility with the original disaggregated constraints.

For problems with special structure a good bound may be found so that fathoming by bounds can be implemented, but for these very general problems, no such good bound can be found easily.

4. COMPUTATIONAL RESULTS FOR LINEARLY CONSTRAINED PROBLEMS

Ten randomly generated problems were solved for each of three sizes of problems with nonlinear objective function and linear constraints. They were solved by Option (a), and results are given in Table 1.

TABLE 1 — *Quadratic Separable Objective Function with Linear Constraints Using Fathoming by Infeasibility*

Problem No.	10 Variables and 5 Constraints	15 Variables and 7 Constraints	20 Variables and 10 Constraints
1	.01	4.31	.02
2	.02	3.86	44.14
3	.01	25.94	16.95
4	.43	79.49	150.37
5	.11	275.73	.03
6	.1	.02	141.63
7	.29	.02	.02
8	.01	.05	36.63
9	.01	.01	.02
10	.01	.01	.05
Average	.1	38.94	38.99
Standard deviation	.15	86.85	58.74

TABLE 2 — *CPU Time in Seconds for Problems With Aggregated Linear Constraints. (x Means Time Limit Exceeded.)*

Problem No.	10 Variables and 5 Constraints	15 Variables and 7 Constraints	20 Variables and 10 Constraints
1	.15	4.05	.06
2	.05	26.91	x
3	.01	x	x
4	1.7	x	x
5	1.89	10.21	.07
6	.36	.09	x
7	4.07	.01	.03
8	.01	.04	x
9	.01	.01	.02
10	.09	.01	17.46
Average	.83	5.17	3.53
Standard deviation	1.34	9.5	7.79

From these results we see that the CPU times on a CDC 6600 in seconds are highly variable. The algorithm was efficient on some of the largest size problems, and comparatively good for medium and small problems. Next we give the computation times for the Option (b) algorithm.

These are identical problems for which the constraints have been aggregated betore fathoming. We see that performance has deteriorated significantly, and many 15 variable problems were not solved to completion. The conclusion is that fathoming is a very important element in the hypersurface search algorithm, and if special structure permits, fathoming by bounding would be expected to improve performance times if it is implemented. Obviously, the tightness of the constraints or bounds is the key factor in evaluating the resulting improvement.

5. NONLINEAR AND NONSEPARABLE CONSTRAINTS

The type of constraint permitted using the algorithm described above is very broad. Nonlinear and even nonseparable constraints are handled without any difficulty in forward pass fathoming. As we generate the forward dynamic programming tables we can look backward at any stage, k, and trace out a partial solution $x_k^*, x_{k-1}^*, \ldots, x_1^*$. Let us assume that the constraint is linear:

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n \leqslant b, \qquad \text{all } a_j \geqslant 0.$$

If we have reached stage k when we do the fathoming test, we are testing the following inequality:

$$\sum_{j=1}^{k} a_j(x_j^*) > b.$$

It is easily seen that if the constraint is nonlinear and *separable*, that is,

$$\sum_{j=1}^{n} h_j(x_j^*) \leqslant b,$$

then fathoming test at the kth stage would be:

$$\sum_{j=1}^{k} h_j(x_j^*) \overset{?}{>} b.$$

Therefore, we have shown that the forward fathoming method can be implemented for one or more nonlinear separable constraints with no additional difficulty in logic or computation.

Now we consider the logic for nonseparable constraints: crossproducts, maximum or minimum expressions, etc. The only change we need to implement in the forward fathoming logic is that a partial solution is not tested for infeasibility until all values x_j^* involved in a nonseparable term are developed. For example, let the constraint be:

$$x_1x_2 + \max\,(x_3, 4x_4) \leqslant b.$$

Instead of testing for infeasibility, that is fathoming at stages 1, 2, 3, and 4, we delay testing until stages 2 and 4. At the second stage the test is

$$x_1^* x_2^* \overset{?}{\leqslant} b.$$

If not, the entry corresponding to $x_2 = x_2^*$ is deleted from the second stage table. No test is made at stage three, but at stage four we test the inequality again with another partial solution:

$$x_1^* x_2^* + \max\,(x_3^*, 4x_4^*) \leqslant b.$$

If the inequality is violated, then the partial solution $(x_1^*, x_2^*, x_3^*, x_4^*)$ is deleted with all its completions. Therefore, the only effect of using the algorithm on nonseparable constraints would be a possible reduction in the efficiency of the fathoming algorithm since it could be implemented only for stages at which all components of a nonseparable term had been developed. An alternative approach, which seems wasteful, would be to use known transformation techniques for factorable functions which result in a separable form, but always at the expense of added variables and constraints.

CONCLUSION

The hypersurface search algorithm has been improved using techniques derived from shortest path problem algorithms in order to allow a reduction of computational effort in evaluating the functional equations of dynamic programming and in reducing the number of entries or states in the tables. We have presented computational results based on these techniques for linear constraints and have shown how they can be used with no additional difficulty for nonlinear and even nonseparable constraints.

REFERENCES

[1] Aris, R., *Discrete Dynamic Programming* (Blaisdale Publishing, New York, 1964).
[2] Baker, K.R., *Introduction to Sequencing and Scheduling* (John Wiley and Sons, Inc., New York, 1974).
[3] Bellman, R., *Dynamic Programming* (Princeton University Press, Princeton, 1957).
[4] Bellman, R., *Adaptive Control Processes* (Princeton University Press, Princeton, 1961).
[5] Bellman, R., *Applied Dynamic Programming* (Princeton University Press, Princeton, 1962).
[6] Bellman, R. and R. Kalaba, "Some Mathematical Aspects of Optimal Prediction in Ecology and Boviculture," *Proceedings, National Academy of Science, 46*, 718, (1960).
[7] Conway, R.W., W.C. Maxwell and L.W. Miller, *Theory of Scheduling* (Addison-Wesley, Reading MA, 1967).
[8] Cooper, L. and M.W. Cooper, *Introduction to Dynamic Programming* (Pergamon, Oxford, Chapter 7, 1981).
[9] Cooper, M.W., "The Use of Dynamic Programming Methodology for the Solution of a Class of Nonlinear Programming Problems," Naval Research Logistics Quarterly *27*, 89-95, (1980).
[10] Cooper, M.W., and K. Farhangian, An Integer Programming Algorithm for Portfolio Selection with Fixed Charges, Naval Research Logistics Quarterly, *28*, 147-150 (1982).
[11] Denardo, E.V. and B.L. Fox, Shortest Route Methods: 1. Reaching, Pruning, and Buckets, Operations Research, *27*, 161-186, (1979).
[12] Gallo, G., P.L. Hammer and B. Simeone, "Quadratic Knapsack Problems," Mathematical Programming Study, *12*, 132-149 (1980).

[13] Glover, F., "Improved Linear Integer Programming Formulations of Nonlinear Integer Problems," Management Science, *22,* 455-460 (1975).
[14] Glover, F. and E. Woolsey, "Further Reduction of Zero-One Polynomial Programming Problems to Zero-One Linear Programming Problems," Operations Research *21,* 141-161 (1973).
[15] Glover, F. and E. Woolsey, "Converting the 0-1 Polynomial Programming Problem to an 0-1 Linear Program," Operations Research, *22,* 180-182 (1974).
[16] Granot, D., F. Granot, and J. Kallberg, "Covering Relaxation for Positive 0-1 Polynomial Programs," Management Science, *25,* 264-273 (1979).
[17] Granot, D. and F. Granot, "Covering Relaxation for 0-1 Programs," Operations Research *28,* 1442-1449 (1980).
[18] Hadley, G. and T.M. Whitin, *Analysis of Inventory Systems,* (Prentice-Hall, Englewood Cliffs, NJ., 1963).
[19] Hammer, P.L. and S. Rudaneau, *Boolean Methods in Operations Research* (Springer-Verlag, New York, 1968).
[20] Hansen, P., "Nonlinear 0-1 Programming by Implicit Enumeration," 7th Mathematical Programming Symposium, The Hague (September, 1971).
[21] Kalaba, R., "Graph Theory and Automatic Control," in *Applied Combinatorial Mathematics* E.F. Beckenbach, Editor. (John Wiley and Sons, New York, 1964).
[22] B. Korte, W.K. Krelle, and W. Oberhofer, "Ein lexikographischer Suchalgorithmus zur Losung allgemeiner ganzzahliger Programmierung-saufgaben," Unternehmensforschung, *13,* 72-98 (1969).
[23] B. Korte, W.K. Krelle, and W. Oberhofer, "Ein lexikographischer Suchalgorithmus zur Losung allgemeiner ganzzahliger Programmierungsaufgaben," Unternehmensforschung, *13,* 171-192 (1969).
[24] B. Korte, W.K. Krelle, and W. Oberhofer, "Ein lexikographischer Suchalgorithmus zur Losung allgemeiner ganzzahliger Programmierung-saufgaben," Unternehmensforschung, *14,* 228-234 (1970).
[25] Laughunn, D.J., "Quadratic Binary Programming with Applications to Capital Budgeting Problems," Operations Research,**18**, 454-461 (1970).
[26] McBride, R.D. and J.S. Yormark, "An Implicit Enumeration Algorithm for Quadratic Integer Programming," Management Science, *26,* 282-296 (1980).
[27] Morin, T.L. and R. Marsten, "Branch and Bound Strategies for Dynamic Programming," Operations Research, *24,* 1147-1158 (1976).
[28] Oron, G., "Optimizing Irrigation Systems Design," Ph.D Dissertation, Technion, Haifa, Israel (1975).
[29] Peterson, D.E. and D. Laughunn, "Capital Expenditure Programming and Some Alternative Approaches to Risk," Management Science, *17,* 320-326 (1971).
[30] Pritsker, A., L.J. Walters, and F. Wolfe, "Multiproject Scheduling with Limited Resources: a Zero-One Programming Approach," Management Science,*16*, 93-108 (1969).
[31] Rao, M.R., "Cluster Analysis and Mathematical Programming," Journal of the American Statistical Association, *66,* 622-626 (1971).
[32] Rhys, J., "A Selection Problem of Shared Fixed Costs and Network Flows," Management Science, *17* 200-207 (1973).
[33] Soland, R.M., "Optimal Defensive Missile Allocation: a Discrete Min-Max Problem," Operations Research, *21,* 590-596 (1973).
[34] Witzgall, C., "An All-Integer Programming Algorithm with Parabolic Constraints," SIAM Journal, *11,* 855-871 (1963).

[35] Witzgall, C., "Mathematical Methods of Site Selection for Electronic Message Systems (EMS)," NBS Internal Report (1975).
[36] Zangwill, W., "Media Selection by Decision Programming," Journal of Advertising Research, *5,* 30-36 (1965).
[37] Zeitlin, Z., "Integer Allocation Problems of Min-Max Type with Quasi-Convex Separable Function," Operations Research, *29,* 207-211 (1981).
[38] Zeitlin, Z., "Some Nonlinear Programming Problems with Integer Resource Allocations," Ph.D. dissertation, Tel-Aviv University, Israel (September 1977).

APPLICATION OF LEVEL CROSSING ANALYSIS TO DISCRETE STATE PROCESSES IN QUEUEING SYSTEMS

J. G. Shanthikumar

Department of Industrial Engineering
and Operations Research
Syracuse University
Syracuse, New York

M. Jeya Chandra

Department of Industrial and
Management Systems Engineering
The Pennsylvania State University
University Park, Pennsylvania

ABSTRACT

In this paper we introduce a discrete state level crossing analysis and present some basic results and a key theorem of level crossings. We illustrate the fertility of the discrete state level crossing analysis by applying it to queueing systems with (i) bulk arrival, (ii) instantaneous feedback, (iii) limited waiting space, and (iv) to machine interference problems.

1. INTRODUCTION

In this paper we define several event epochs associated with a general queueing system and derive relations among the steady state probabilities associated with these event epochs. Often, in queueing theory, embedded Markov chains are used to obtain the steady state probability distributions of the number in the system at arrival or departure epochs. Then specific approaches, suited to individual queueing problems, are used to relate the steady state probability distributions associated with different event epochs such as arrival, departure, feedback, and rejection and with arbitrary time epoch. For example, see Chaudhry [7], Chow [8], Courtois and Georges [14], Disney, McNickle and Simon [15], Heyman and Stidham [18], Neuts [22], and Takacs [33].

In this paper we present some basic relations among these steady state probabilities and a key theorem of level crossings relating the steady state probabilities at arrival and departure epochs. The fundamental idea behind this key theorem is the relationship between the up and downcrossings of regenerative processes.

Level crossing analysis that utilizes the relationship between up and down crossings of stochastic processes with continuous state space is introduced by Brill [1], Brill and Posner [3,4,5],

and independently by Cohen [9,10]. Also an alternate imbedded level crossing analysis is discussed by Brill [2]. Shanthikumar [23] extended Cohen's [9,10] approach to alternating regenerative processes with continuous state space and illustrated its applicability to several queueing problems with server interruptions and several classes of customers (see [24-27]). Shanthikumar [28] also introduced an alternate form of level crossing analysis to regenerative processes with continuous state space.

All the approaches mentioned above are specifically designed for stochastic processes with continuous state space and do not directly extend themselves to stochastic process with discrete state space. In this regard, we introduce a new level crossing analysis that is applicable to stochastic processes with discrete state space. The fertility of this method is illustrated by a simpler derivation of existing results for some queues with bulk arrival and customer feedbacks. New results are derived for general queueing systems and a machine interference problem. Further application of this approach to develop new and efficient algorithmic solutions to a computer system model is illustrated in [30] and to an automatic transfer line model in [32]. Also, using this discrete state level crossing analysis a new hybrid simulation/analytic model is developed for a computerized manufacturing system (see [31]). See Shanthikumar [29] for a recent application of this method to a single server queue with server failures and state-dependent arrival process. In the past, some basic exploitation of this idea of discrete state level crossing analysis has been carried out by Cooper [11], Gross and Harris [17], and Jain and Seth [19].

In Section 2 we present some basic relations and the key theorem of level crossings. We illustrate the fertility of this level crossing analysis, in Section 3, by applying it to queueing systems with (i) bulk arrival, (ii) instantaneous feedback, (iii) limited waiting space, and to (iv) machine interference problems.

2. THE BASIC RESULTS

In this section, we will derive some basic results and a key theorem of level crossings for a special case of regenerative process, which occur in many queueing systems. These results will be used later to analyze some queueing systems.

The Regenerative Process

A regenerative process $\{X(t),\ t \geqslant 0\}$ is a stochastic process which starts anew probabilistically at an increasing sequence $0 \leqslant R_1 < R_2 < \ldots$ of random epochs on the time axis $[0, \infty)$. Thus, between any two consecutive regenerative epochs R_i and R_{i+1}, the portion $\{X(t),\ R_i \leqslant t < R_{i+1}\}$ called regeneration cycle, is an independent and identically distributed replicate of the portion between any other two consecutive regeneration epochs [9,13]. We represent the time intervals between any consective regeneration epochs by the set $\{T_i,\ i \geqslant 1\}$ where $T_i = R_{i+1} - R_i$, and assume $E(T_i) < \infty$ for all values of $i \geqslant 1$.

The process $\{X(t),\ t \geqslant 0\}$ is assumed to have a state space N^+ of nonnegative integers and to have sample paths which are continuous from the right and which make only a finite number of jumps in each finite time interval. In almost all queueing systems, the epochs when customers arrive to find *all* servers idle constitutes a sequence of regeneration epochs. If $X(t)$ denotes the number of customers at time t, $t \geqslant 0$, then a sample realization of this process can be represented by Figure 1. In the queueing context, an accepted external arrival of a batch of

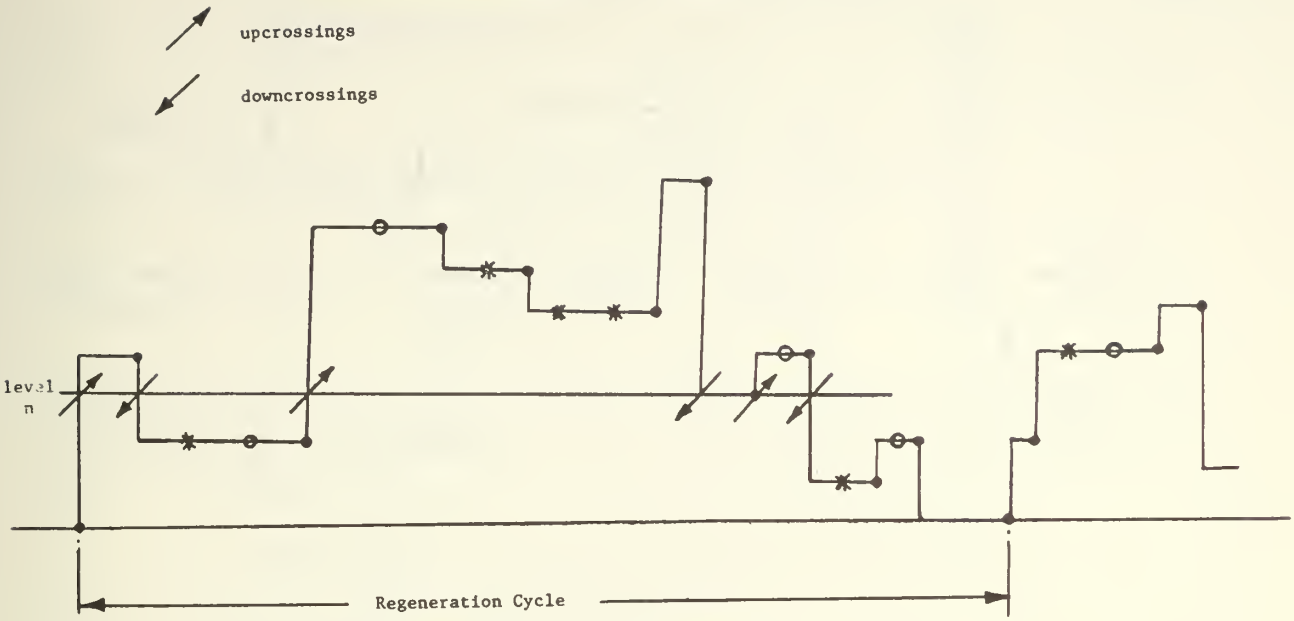


FIGURE 1. A sample realization of the regenerative process $X(t)$

customers at the system is represented by an upward jump and a departure from the system by a downward jump. The magnitudes of the upward and downward jumps are assumed to be independent and depend only on the level (that is, the state) of $X(t)$ just before the upward and downward jumps, respectively. The magnitudes of these upward and downward jumps from level n are assumed to have distributions $\{g_n(k),\ k \geqslant 1\}$ and $\{f_n(k),\ k \geqslant 1\}$ with finite means $E(X_n)$ and $E(Y_n)$, respectively. We allow for the possibility of rejections. That is, an external arrival of a batch of customers seeing a level n may be rejected (that is, entry to the system is denied) with probability g_n. These epochs of rejections are marked by stars in the sample path. Similarly, feedback is also allowed (see Figure 2). That is, customers, after service completion, rejoin the queue with probability f_n, for more service instead of leaving the system. These epochs are represented by hollow circles in the sample path. Note that neither an upward nor a downward jump occurs at the rejection or feedback epoch.

Let the sets U, D, R and F contain the epochs of upward jumps, downward jumps, rejections and feedbacks, respectively, within the first regeneration cycle during $[R_1, R_2)$. Then $A = U \cup R$, $C = D \cup F$, and $I = U \cup F$ are the sets of external arrival epochs, service completion epochs and input epochs to the system including feedback, respectively. Now for each level n of $X(t)$, define U_n^-, D_n^-, R_n^-, F_n^-, A_n^-, C_n^-, and I_n^- as the sets of epochs of upward jumps from level n, downward jumps from level n, rejections at level n, feedback at level n, arrivals to see level n, service completions at level n, and inputs to see level n, respectively, during the first regeneration cycle.

Similarly define U_n^+, D_n^+, R_n^+, F_n^+, A_n^+, C_n^+, and I_n^+ as the sets of epochs of upward jumps to level n, downward jumps to level n rejections at level n, feedback at level n after arrivals, service completions to level n, and level n after inputs, respectively, during the first

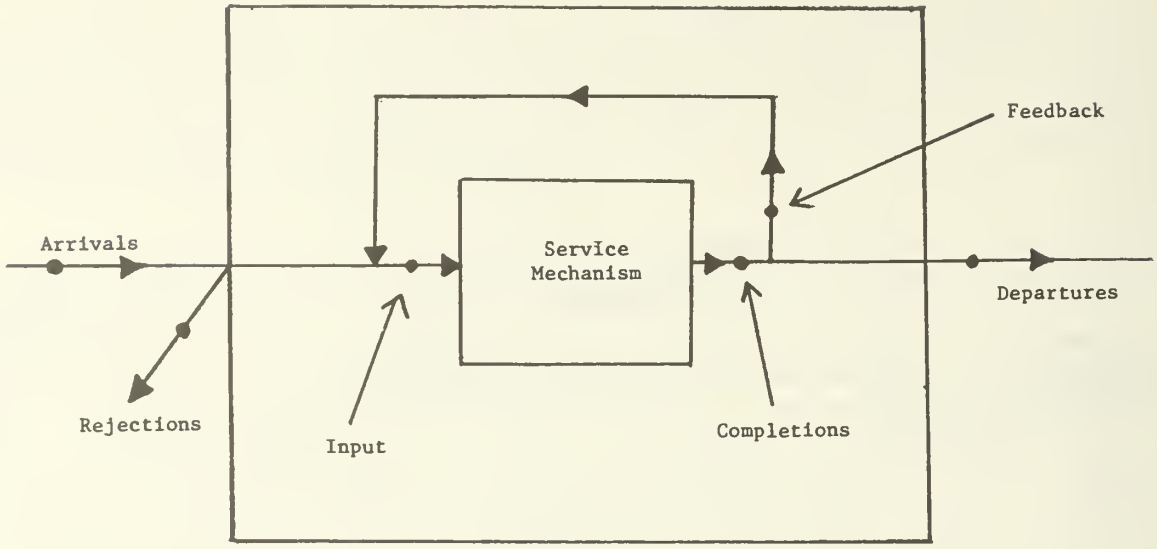


FIGURE 2. Overall system

regeneration cycle. The sets E_n^- and E_n^+ basically represent the epochs at which the level is n just before and just after the occurrence of the event e, for $e \in \{u, d, r, f, a, c, i\} \triangleq S$, generically associated with the set $E \in \{U, D, R, F, A, C, I\}$ of events as defined above. Note that $R_n^- \equiv R_n^+$, and $F_n^- \equiv F_n^+$.

Let ${}_eT_k$ be the time of the kth occurrence of event e. Then if $\#\{E\}$ is the cardinality of a set E and $\delta(\cdot)$ is the indicator function, we have

$$\tau^e \triangleq \#\{k: R_1 \leqslant {}_eT_k < R_2\} = \#\{E\}$$

$${}_e\tau_n^- \triangleq \sum_{k=1}^{\tau^e} \delta(X({}_eT_k^-) = n) = \#\{E_n^-\},$$

and

$${}_e\tau_n^+ \triangleq \sum_{k=1}^{\tau^e} \delta(X({}_eT_k^+) = n) = \#\{E_n^+\},$$

when e and E are generically matched. Now we will define the associated probability distributions (see Figure 2)

$$p_e^-(n) = \lim_{k\to\infty} P\{X({}_eT_k^-) = n\}$$

and

$$p_e^+(n) = \lim_{k\to\infty} P\{X({}_eT_k^+) = n\}.$$

The existence of these distributions is guaranteed by the regenerative property of $\{X(t),\ t \geqslant 0\}$. From the mean value representation for the regenerative processes, we have

$$p_e^-(n) = \frac{E\{{}_e\tau_n^-\}}{E\{\tau^e\}}$$

and

$$p_e^+(n) = \frac{E\{{}_e\tau_n^+\}}{E\{\tau^e\}}$$

Based on the earlier observation that $R_n^- \equiv R_n^+$ and $F_n^- \equiv F_n^+$, one has

$$p_f^-(n) = p_f^+(n) \triangleq p_f(n), \quad n \geq 1$$

and

$$p_r^-(n) = p_r^+(n) \triangleq p_r(n), \quad n \geq 0.$$

Relationships between $p_e^-(n)$ and $p_e^+(n)$ are an immediate consequence of the above "stochastic mean value" equations. For example, we have the following:

(1) $$p_a^+(n) = \sum_{i=0}^{n-1} p_a^-(i)\, g_i(n-i)(1-g_i) + p_a^-(n) g_n, \quad n \geq 0$$

(2) $$p_u^+(n) = \sum_{i=0}^{n-1} p_u^-(i)\, g_i(n-i), \quad n > 0$$

(3) $$p_r^+(n) = p_r^-(n) = p_r(n), \quad n \geq 0$$

(4) $$p_f^+(n) = p_f^-(n) = p_f(n), \quad n \geq 1$$

(5) $$p_d^+(n) = \sum_{i=n+1}^{\infty} p_d^-(i)\, f_i(i-n) \quad n \geq 0$$

$$p_c^+(n) = \sum_{i=n+1}^{\infty} p_c^-(i)\, f_i(i-n)(1-f_i) + p_c^-(n) f_n, \quad n \geq 1$$

(6) $$= \sum_{i=1}^{\infty} p_c^-(i)\, f_i(i)(1-f_i), \quad n = 0$$

Let r_e be the rate of occurrence of events of type e. That is,

$$r_e = E\{\tau^e\}/E\{T_1\}, \quad \forall\, e \in S.$$

Using the relationship between the rates of occurrence of different event epochs, we will derive the following relations:

(7) $$p_u^-(n) = p_a^-(n)(1-g_n)/\left[1 - \sum_{n=0}^{\infty} p_a^-(n) g_n\right], \quad n \geq 0$$

(8) $$p_r(n) = p_a^-(n) g_n/\left[\sum_{n=0}^{\infty} p_a^-(n) g_n\right], \quad n \geq 0$$

(9) $$p_d^-(n) = p_c^-(n)(1-f_n)/\left[1 - \sum_{n=1}^{\infty} p_c^-(n) f_n\right], \quad n > 0$$

(10) $$p_f(n) = p_c^-(n) f_n/\left[\sum_{n=1}^{\infty} p_c^-(n) f_n\right], \quad n > 0$$

Let the sets U_n^-, D_n^-, R_n^-, F_n^-, A_n^-, C_n^-, and I_n^- be as defined earlier. Now define (generically subscripting) $r_u^-(n)$, $r_d^-(n)$, $r_r^-(n)$, $r_f^-(n)$, $r_a^-(n)$, $r_c^-(n)$, and $r_i^-(n)$ as the rates of upward jumps, departure, rejections, feedbacks, arrivals, completions, and inputs with respect to level n of the $X(t)$ process, such that,

$$r_e^-(n) = E({}_e\tau_n^-)/E(T_1), \quad n \geqslant 0, \tag{11}$$

for $e \in S \triangleq \{u, d, r, f, a, c, i\}$ and $E \in \{U, D, R, F, A, C, I\}$, where the pair (e,E) is chosen to match generically. Note that

$$r_e^-(0) = 0 \text{ for } e \in \{f, c, d\}$$

and, since $\overset{\infty}{\underset{n=0}{U}} E_n^- = E$,

$$r_e = \sum_{n=0}^{\infty} r_e^-(n), \ \forall\, e \in S. \tag{12}$$

Then from the "stochastic mean value equation" for $p_e^-(n)$, we have

$$p_e^-(n) = r_e^-(n)/r_e, \ \forall\, e \in S. \tag{13}$$

Therefore, the interrelationships between these probabilities can be derived using the relationship between the rates $r_e^-(n)$, $e \in S$. Since the probability of rejection of a batch on its arrival is g_n if it sees n in the system, we have

$$r_u^-(n) = r_a^-(n)(1 - g_n), \quad n \geqslant 0 \tag{14}$$

and

$$r_r^-(n) = r_a^-(n)g_n, \quad n \geqslant 0. \tag{15}$$

Then from (12), (13), and (14), equation (7) follows. Similarly, from (12), (13), and (15), equation (8) follows. A similar inspection of completion epochs gives

$$r_d^-(n) = r_c^-(n)(1 - f_n), \quad n > 0 \tag{16}$$

and

$$r_f^-(n) = r_c^-(n)f_n, \quad n > 0. \tag{17}$$

Then similar to the above cases, equations (9) and (10) follow from (12), (13), (16), and (17).

Note that when $g_n = g$, $\forall\, n \geqslant 0$, from (7) and (8) $p_a^-(n) = p_u^-(n) = p_r(n)$, and similarly when the feedback probability f_n is independent of n, that is, $f_n = f, \forall\, n \geqslant 1$, from (9) and (10), $p_c^-(n) = p_d^-(n) = p_f(n)$.

Next we will establish the following relations:

$$p_c^+(n) = \frac{p_d^+(n) + \left(\dfrac{f_n}{1 - f_n}\right) p_d^-(n)}{1 + \displaystyle\sum_{n=1}^{\infty} \left(\frac{f_n}{1 - f_n}\right) p_d^-(n)}, \quad n \geqslant 1 \tag{18}$$

$$p_i^-(n) = \frac{p_u^-(n)r_u + r_d\left(\frac{f_n}{1-f_n}\right)p_d^-(n)}{r_u + r_d\sum_{n=1}^{\infty}\left(\frac{f_n}{1-f_n}\right)p_d^-(n)}, \quad n \geqslant 1 \tag{19}$$

$$p_i^+(n) = \frac{p_u^+(n)r_u + r_d\left(\frac{f_n}{1-f_n}\right)p_d^-(n)}{r_u + r_d\sum_{n=1}^{\infty}\left(\frac{f_n}{1-f_n}\right)p_d^-(n)}, \quad n \geqslant 1 \tag{20}$$

where r_u and r_d are the rates of upward and downward jumps as defined earlier.

Now as defined earlier, let U_n^+, I_n^+, D_n^+, and C_n^+ be the sets consisting of epochs, just after an upward jump, input, downward jump, and completion when $X(t) = n$. That is,

$$E_n^+ = \{t: X(t^+) = n,\ t \in E\},$$

for

$$E \in \{U,\ I,\ D,\ C\}.$$

Now generically define the rates

$$r_e^+(n) = E\{{}_e\tau_n^+\}/E\{T_1\},\ e \in \{u,\ i,\ d,\ c\},\ n \geqslant 0 \tag{21}$$

and since $\bigcup_{n=0}^{\infty} E_n^+ \equiv E$, we have

$$r_e = \sum_{n=0}^{\infty} r_e^+(n).$$

Then, as in the previous case.

$$p_e^+(n) = r_e^+(n)/r_e,\ e \in \{u,\ i,\ d,\ c\}. \tag{22}$$

Note that

$$r_c^+(n) = r_f(n) + r_d^+(n), \tag{23}$$

and therefore, from (21), (22), and (23) we get equation (18), after noting from (16) and (17) that $r_f(n) = r_d^-(n)f_n/(1-f_n)$. Similarly, the relation

$$r_i^-(n) = r_u^-(n) + r_f(n) \tag{24}$$

and

$$r_i^+(n) = r_u^+(n) + r_f(n) \tag{25}$$

along with $r_f(n) = r_d^-(n)f_n/(1-f_n)$, (21) and (22) lead to (19) and (20).

Equating the number of customers arrived and departed during a regeneration cycle, we get

$$\sum_{n=0}^{\infty} r_u^-(n)E(X_n) = \sum_{n=1}^{\infty} r_d^-(n)E(Y_n). \tag{26}$$

These relations (1), (2), (3), (7), and (8) form one class such that if $p_a^-(n)$, $n \geq 0$, is known, we can derive $p_a^+(n)$, $p_u^-(n)$, $p_u^+(n)$, and $p_r(n)$. Similarly, the relations (4), (5), (6), (9), (10), and (18) form a second class such that if $p_d^-(n)$, $n \geq 1$, is known we can obtain $p_d^+(n)$, $p_f(n)$, $p_c^-(n)$, and $p_c^+(n)$. However, to relate these two sets of probabilities and to use (19) and (20), we need a relation between two probability distributions, one from each set. Such a relation is our key theorem of level crossings, which we present next.

THEOREM 1: The Key Theorem of Level Crossing

The probability distributions $p_u^-(n)$ and $p_d^-(n)$ are related to one another by:

$$r_u \sum_{i=0}^{n} p_u^-(i)(1 - G_i(n-i)) = r_d \sum_{i=n+1}^{\infty} p_d^-(i)(1 - F_i(i-n-1)), \quad n \geq 0 \tag{27}$$

where

$$G_i(k) = \sum_{j=1}^{k} g_i(j), \; G_i(0) = 0 \text{ and } F_i(k) = \sum_{j=1}^{k} f_i(j), \; F_i(0) = 0$$

are the cumulative distributions of $g_i(\cdot)$ and $f_i(\cdot)$, respectively.

PROOF: Define

$\# U_n$ = $\#\{t: X(t^-) \leq n, X(t^+) > n, t \in U\}$, the number of upcrossings over level n from and below level n, and

$\# D_n$ = $\#\{t: X(t^-) \geq n+1, X(t^+) < n+1; t \in D\}$, the number of downcrossings below level $n+1$ from and above level $n+1$, by the process $\{X(t), t \geq 0\}$ during its first regeneration cycle.

A simple inspection of the sample path (see Figure 1) will show that

$$\# U_n = \# D_n \text{ a.s.} \tag{28}$$

From the definition of $\# U_n$, it is clear that

$$E\{\# U_n\} = E\left\{\sum_{k=0}^{\tau^u} \delta(X({}_uT_k^-) \leq n < X({}_uT_k^-) + X^k)\right\},$$

where X^k is the batch size at the k-th upcrossing in the first regeneration cycle. Since

$$\delta(X({}_uT_k^-) \leq n < X({}_uT_k^-) + X^k) = \delta(X({}_uT_k^-) \leq n) - \delta(X({}_uT_k^-) + X^k \leq n),$$

from the above equation for $E\{\# U_n\}$, we get

$$E\{\# U_n\} = E\left\{\sum_{k=0}^{\tau^u} \delta(X({}_uT_k^-) \leq n) - \sum_{k=0}^{\tau^u} \delta(X({}_uT_k^-) + X^k \leq n)\right\}.$$

Now from the "mean value equations" for $p_e^-(n)$, $p_e^+(n)$, and from (2), we have

$$E\{\# U_n\} = E\{\tau^u\} \sum_{i=0}^{n} p_u^-(i)(1 - G_i(n-i)). \tag{29}$$

Using a similar derivation, we can show that

$$E\{\#D_n\} = E\{\tau^d\} \sum_{i=n+1}^{\infty} p_d^-(i)(1 - F_i(i - n - 1)). \tag{30}$$

Now from (28), equating (29) and (30) and dividing by $E(T_1)$, we get (27).

COROLLARIES:

Next for a special case of the $X(t)$ process we will relate the up and downcrossing rates to the steady state probability $q(\cdot)$ at an arbitrary time. The special case considered is a piecewise Markov process (see Kuczura [21]). For this, we assume that there exists a sequence of random variables $\{W_n\}_0^\infty$ with $0 < W_1 < W_2 < \ldots$, such that for any $n \geqslant 1$,

$$(i)\ \{X(t),\ W_{n-1} \leqslant t < W_n\},$$

with $W_0 = 0$, is a Markov process

$$(ii)\ p_{ij} = P\{X(W_n) = j | X(W_n^-) = i\},$$

is the same for all $n \geqslant 1$, and

$$(iii)\ Q_k(u) = P\{W_n - W_{n-1} \leqslant u | X(W_{n-1}) = k\},$$

is a family of distribution functions of positive random variables.

CASE 1: Let us first assume that the process $\{X(t),\ t \geqslant 0\}$ is piecewise Markovian between two successive downward jumps such that,

$$P\{\text{upward jump during}(t,\ t + \Delta t) | X(t) = n\} = \lambda(n)\Delta t + 0(\Delta t),$$

$\forall\ n \geqslant 0$, where $\lim_{\Delta t \to \infty} \dfrac{0(\Delta t)}{\Delta t} = 0$, is independent of time t and depends only on the state of $X(t)$. From equation (27) and the rate of upcrossings, we then have

COROLLARY 1:

$$\sum_{i=0}^{n} \lambda(i)q(i)(1 - G_i(n - i)) = r_u \sum_{i=0}^{n} p_u^-(i)(1 - G_i(n - i))$$

$$= r_d \sum_{i=n=1}^{\infty} p_d^-(i)(1 - F_i(i - n - 1)) \tag{31}$$

CASE 2. For this case we will assume that $X(t)$ is a piecewise Markov process between successive upward jumps so that,

$$P\{\text{downward jump during}(t,\ t + \Delta t) | X(t) = n\} = \mu(n)\Delta t + 0(\Delta t),$$

$n \geqslant 1$, is independent of t and depends only on $X(t)$. Similar to Case 1,

COROLLARY 2:

$$\sum_{i=n+1}^{\infty} \mu(i)q(i)(1 - F_i(i - n - 1)) = r_u \sum_{i=0}^{n} p_u^-(i)(1 - G_i(n - i))$$

$$= r_d \sum_{i=n+1}^{\infty} p_d^-(i)(1 - F_i(i - n -)) \tag{32}$$

In the next section we will illustrate the fertility of these results by applying them to several queueing examples.

3. APPLICATIONS

3.1 Queueing Systems With Bulk Arrival

Let us consider a queueing system to which the arrivals are in batches of size B. Let $P(B = n) = b_n$, $n \geqslant 1$. It is assumed that the interarrival times from a renewal sequence with mean $1/\lambda$. The departures, however, occur singly such that the magnitude of a downward jump is one. It is assumed that all arrivals are accepted. However, no specific assumptions are made regarding the feedback. A special case of this system is a $GI^x/G/c$ queue with the restriction that departures occur singly. Now from (26) and (12), we get

$$r_u = \lambda \text{ and } r_d = \lambda E(B) \tag{33}$$

Now using (33) in (27), after noting that $p_a^-(n) = p_u^-(n)$ and $p_d^-(n+1) = p_d^+(n)$,

$$\sum_{i=0}^{n} p_a^-(i)(1 - H(n-i)) = E(B)p_d^+(n), \quad n \geqslant 0, \tag{34}$$

where

$$H(k) = \sum_{j=1}^{k} b_j = P(B \leqslant k), \quad H(0) = 0.$$

Now taking the moment generating function of (34) we get

$$\frac{\tilde{P}_a^-(z)(1 - \tilde{H}(z))}{(1 - z)} = E(B)\tilde{P}_d^+(Z), \tag{35}$$

where

$$\tilde{P}_a^- = \sum_{n=0}^{\infty} z^n p_z^-(n), \quad \tilde{P}_d^+(z) = \sum_{n=0}^{\infty} z^n p_d^+(n),$$

and

$$\tilde{H}(z) = \sum_{n=1}^{\infty} z^n b_n$$

are the moment generating functions of $p_a^-(\cdot)$, $p_d^+(\cdot)$ and b_n. This result relating the steady state probability distribution seen by an arriving batch and a departing customer immediately extends the key result (equation (1)) derived by Chaudhry [7] for an $M^x/G/1$ queue to systems even more general than a $GI^x/G/c$ feedback queue. Now let us consider two special cases of the above model: (i) $GI^x/M/c$ and (ii) $M^x/G/c$.

3.1.1 $GI^x/M/c$ Queue

Let the service times be exponentially distributed with mean $1/\mu$ and the feedback probability $f_n = f$, $n \geqslant 1$. Then from corollary 2, with $\mu(n) = \min(n,c)\mu(1-f)$, and (34), we get

$$\lambda \sum_{i=0}^{n} p_a^-(i)(1 - H(n-i)) = \min(n+1,c)\mu(1-f)q(n+1) \tag{36}$$

Result (36) is a generalization of the results given in Takacs [33] for the $GI/M/c$ gueue.

3.1.2 $M^x/G/c$ Queue

Here we assume that the arrivals of batches form a Poisson process with rate λ. Then from Corollary 1 we have $q(n) = p_a^-(n)$ and from (35) we get

$$\tilde{Q}(z) = \tilde{P}_a^-(z) = \{E(B)(1 - z)\}\tilde{P}_d^+(z)/(1 - \tilde{H}(z)), \tag{37}$$

where $\tilde{Q}(z) = \sum_{n=0}^{\infty} z^n q(n)$, and others are as defined earlier. Result (37) was derived by Chaudhry [17] (Equation (1)) for the $M^x/G/1$ queue.

Similar results can be derived for the bulk service queues.

3.2 Queues With Instantaneous Feedback

Now we will consider a queueing system with single arrivals and departures. That is, $g_n(1) = f_n(1) = 1\ \forall n$. We will also assume that the probability of blocking g_n is zero and probability of feedback f_n is a constant equal to f, $\forall n \geqslant 1$. A special case of this system is the $GI/G/c$ queue with instantaneous feedback. Let the mean inter-arrival time be $1/\lambda$. Recently Disney, McNickle and Simon [15] considered the $M/G/1$ queue with instantaneous feedback. They obtained relationships between the probabilities at different event epochs using Markov Renewal theory. Our result will generalize theirs. From (26) and (12), we have $r_u = r_d$. Now from the key theorem of level crossing, and noting that $p_a^-(n) = p_u^-(n)$,

$$p_a^-(n) = p_d^-(n + 1) = p_d^+(n),\ n \geqslant 0, \tag{38}$$

Now from (18) and (38),

$$p_c^+(n) = (1 - f)p_d^+(n) + f\, p_d^+(n - 1),\ n \geqslant 1. \tag{39}$$

From (4), (9), and (10),

$$p_f(n) = p_c^-(n) = p_d^-(n),\ n \geqslant 1. \tag{40}$$

Also from (6),

$$\left.\begin{aligned} p_c^+(n) &= p_c^-(n + 1)(1 - f) + p_c^-(n)f,\ n \geqslant 1 \\ &= p_c^-(1)(1 - f),\ n = 0 \end{aligned}\right\} \tag{41}$$

So having evaluated $p_d^+(\cdot)$ from embedded Markov chain analyses, the remaining probability distributions can be evaluated using (38)-(41).

3.3 Queueing Systems With Limited Waiting Space

In this section we consider the same model as 3.2 except that the capacity of the system is K. That is, $g_n = 1$, $n \geqslant K$ and $= 0$, $n < K$. Then from (7) we get

$$p_u(n) = p_a^-(n)/(1 - p_a^-(K)),\ 0 \leqslant n < K \tag{42}$$

and

$$p_a^-(n + 1) = p_u^-(n) = 0,\ n \geqslant K.$$

Now from the key theorem of level crossings and noting that $r_u = r_d$,

$$p_u^-(n) = p_d^-(n + 1),\ 0 \leqslant n < K. \tag{43}$$

Next we will consider two special cases.

3.3.1 GI/M/c/K Queue

From Corollary 2 we get, after noting that $\mu(n) = \min(n,c)\mu(1-f),\ n = 1, 2, \ldots, K$,

$$r_u p_u^-(n-1) = \min(n,c)\mu(1-f)q(n),\ n = 1, 2, , \ldots, K. \tag{44}$$

From (42) and noting that $r_u = \lambda(1 - p_a^-(K))$, we get

$$\lambda p_a^-(n-1) = \min(n,c)\mu(1-f)q(n),\ n = 1,2, \ldots, K. \tag{45}$$

The above result agrees with that of Heyman and Stidham [18].

3.3.2 M/G/1/K Queue

From (42), (43), and Corollary 1 we get

$$\begin{aligned} \lambda q(n) &= r_u p_u^-(n),\ n = 0, 1, 2, \ldots, K-1, \\ &= r_d p_d^-(n+1). \end{aligned} \tag{46}$$

Now noting that $r_d = \mu(1 - q(0))$, we get from (46)

$$q(n) = \frac{(1-q(0))p_d^-(n+1)}{\rho},\ n = 0, 1, 2, \ldots, K-1, \tag{47}$$

where $\rho = \lambda/\mu$. So

$$q(0) = \frac{p_d^+(0)}{\rho + p_d^+(0)},$$

$$q(n) = \left(\frac{1}{\rho + p_d^+(0)}\right) p_d^+(n),\ n = 1, 2, \ldots, K-1, \tag{48}$$

and

$$q(K) = 1 - \sum_{n=0}^{K-1} q(n)$$

give the steady state probability distribution at an arbitrary time epoch. See Gross and Harris [17, page 252], for an alternate derivation of the same results.

4. MACHINE INTERFERENCE PROBLEMS

In this section we apply the level crossing technique to machine interference models which are queueing models with finite capacity sources, and derive some new results for this model.

4.1 M/G/1/ /N Queue

We consider here a queueing model with a single finite source model of capacity N and a single server. The arrival rate of customers, when there are n customers at the service facility, is $\lambda(N-n)$. Besides, $g_n = 1$ when $n \geq N$ and 0 when $n < N$, and $p_a^-(n) = 0$ if $n \geq N$. So we get from (7)

$$\begin{aligned} p_u^-(n) &= p_a^-(n),\ 0 \leq n < N \\ &= 0 \qquad \text{Otherwise,} \end{aligned} \tag{49}$$

from (26), $r_u = r_d$.

From Corollary 1, and because of (49), we get for $0 \leqslant n < N$

$$\lambda(n)\, q(n) = r_u p_u^-(n) = r_d\, p_d^-(n+1). \tag{50}$$

Let $1/\mu$ represent the mean service time of a customer, and ρ the utilization of the server. Then,

$$r_d = \mu\rho.$$

Because $p_d^-(n+1) = p_d^+(n)$, we get from (50), for $0 \leqslant n < N$,

$$\lambda(n)\, q(n) = \mu\rho p_d^+(n).$$

So

$$q(n) = \frac{\mu\rho\ p_d^+(n)}{\lambda(n)} \quad \text{for } 0 \leqslant n < N$$

and

$$q(N) = 1 - \sum_{n=0}^{N-1} q(n). \tag{51}$$

These results agree with equation (15) of Courtois and Georges [14] and equation (10) of Chow [8].

4.2 M/G/1/ /N_1, N_2, ... N_k Queue

We now extend the results to a multiple finite source queueing model with K finite sources and a single server. The capacity of the ith source is N_i. The server can follow any work-conserving service discipline for selecting customers for service. Relations are obtained in the following subsections with respect to marginal and joint probabilities.

4.2.1 Marginal Probabilities

We consider each source separately and apply results obtained in Section 4.1. We define

$$p_d^{k+}(n_k) = \lim_{r\to\infty} P(X_k(\xi_r^+) = n_k)$$

where $X_k(t)$ denotes the number of source k customers present at time t, $t \geqslant 0$, and ξ_r is the service completion epoch of the rth type k customer. So from (51) we get, for $k = 1, 2, \ldots, K$,

$$q^k(n_k) = \frac{\mu_k \rho_k}{\lambda_k(n_k)} \cdot p_d^{k+}(n_k), \text{ if } 0 \leqslant n_k < N_k$$

and

$$q^k(N_k) = 1 - \sum_{i=0}^{N_k - 1} q^k(i) \tag{52}$$

where $q^k(n_k)$ is the marginal time average probability of finding n_k number of source k customers at the facility, ρ_k is the proportion of time the server is busy with source k customers and $1/\mu_k$ is the mean service time of a source k customer. Then $r_d = \mu_k \rho_k$ for class k source.

4.2.2 Joint Probabilities

In this case, we represent the state of the system $X(t)$ as a vector $\underline{n} = (n_1, n_2, \ldots, n_K)$, where n_k refers to the number of source k customers.

From an extension of Corollary 1, for the multidimensional discrete state processes, we can show that, for the compound state $\{\mathbf{i};\ i_j \leqslant n_j,\ j = 1, 2, \ldots, K\}$ with $\sum_{j=1}^{K} n_j \neq \sum_{j=1}^{K} N_j$,

$$\sum_{k=1}^{K} \lambda_k(n_k) \left\{ \sum_{i_1=0}^{n_1} \sum_{i_2=0}^{n_2} \cdots \sum_{i_{k-1}=0}^{n_{k-1}} \sum_{i_{k+1}=0}^{n_{k+1}} \cdots \sum_{i_K=0}^{n_K} q(i_1, i_2, \ldots, i_{k-1}, n_k, i_{k+1}, \ldots, i_K) \right\}$$

$$= \sum_{k=1}^{K} \mu_k \rho_k \left\{ \sum_{i_1=0}^{n_1} \sum_{i_2=0}^{n_2} \cdots \sum_{i_{k-1}=0}^{n_{k-1}} \sum_{i_{k+1}=0}^{n_{k+1}} \cdots \sum_{i_K=0}^{n_K} p_d^{k+}(i_1, i_2, \ldots, i_{k-1}, n_k, i_{k+1}, \ldots, i_K) \right\} \tag{53}$$

where $p_d^{k+}(\mathbf{n}) = \lim_{r \to \infty} P(X(\xi_r^+) = \mathbf{n})$.

Rearranging the terms, (53) becomes

$$q(\underline{n})\lambda(\underline{n}) = \sum_{k=1}^{K} p_d^{k+}(\mathbf{n})\mu_k \rho_k - \sum_{k=1}^{K}$$

$$\left\{ \sum_{i_1=0}^{n_1} \sum_{i_2=0}^{n_2} \cdots \sum_{i_{k-1}}^{n_{k-1}} \sum_{i_{k+1}=0}^{n_{k+1}} \cdots \sum_{i_K=0}^{n_K} (q(i_1, i_2, \ldots, i_{k-1}, n_k, i_{k+1}, \ldots, i_K)\lambda_k(n_k) \right.$$

$$\left. - p_d^{k+}(i_1, i_2, \ldots, i_{k-1}, n_k, i_{k+1}, \ldots, i_K)\mu_k \rho_k) \right\} \tag{54}$$

$$\sum_{j=1}^{K} i_j \neq \sum_{j=1}^{K} n_j$$

$$q(N_1, N_2, \ldots, N_k) = 1 - \left\{ \sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \cdots \sum_{n_k=0}^{N_k} q(n_1, n_2, \ldots, n_K) \right\}$$

$$\sum_{k=1}^{K} n_i \neq \sum_{k=1}^{K} N_k. \tag{55}$$

Details of the numerical algorithms using the above relations for calculating the time average probabilities $q(\underline{n})$ are given in [6]. A detailed derivation of these and other results with numerical examples will be reported elsewhere.

5. CONCLUSIONS

In this paper we have presented some basic results concerning level crossing analysis applied to discrete state processes. By applying these results to some queueing systems, we obtained relations between time average probabilities and the steady state probabilities at arrival or departure epochs.

The queueing systems considered in this paper are but a few among the vast number of cases where the level crossing analysis can be used to relate the steady probabilities associated with different event epochs. The most important feature of this analysis is that the level crossing analysis is a more general and simple tool which can be applied to many queueing systems. Application of this discrete state level crossing analysis has already proven to be very fruitful in different cases (see [29-32]).

ACKNOWLEDGMENT

The authors would like to thank the referee and Dr. Daniel P. Heyman, Bell Labs for their valuable suggestions. They also pointed out a technical error in an earlier version of this paper and brought some references to the authors' attention.

BIBLIOGRAPHY

[1] Brill, P.H., "System Point Theory in Exponential Queues," Ph.D. Thesis, Department of Industrial Engineering, University of Toronto, Toronto, Canada (1975).

[2] Brill, P.H., "An Embedded Level Crossing Technique for Dams and Queues," Journal of Applied Probability, *16*, 174-186 (1979).

[3] Brill, P.H. and M.J.M. Posner, "Level Crossing in Point Processes Applied to Queues: Single Server Case," Operations Research, *25*, 662-674 (1977).

[4] Brill, P.H. and M.J.M. Posner, "The System Point Method in Exponential Queues: A Level Crossing Approach," Mathematics of Operations Research, *6*, 31-49 (1981).

[5] Brill, P.H. and M.J.M. Posner, "A Two Server Queue with Nonwaiting Customers Receiving Specialized Service," Management Science, *27*, 914-925 (1981).

[6] Chandra, M.J., "Multiple Finite Source Model With Fixed Priority Scheduling," Ph.D. Dissertation, Department of Industrial Engineering and Operations Research, Syracuse University, Syracuse, NY (1980).

[7] Chaudhry, M.L., "The Queueing System $M^x/G/1$ and Its Ramifications," Naval Research Logistics Quarterly, *26*, 667-674 (1979).

[8] Chow, W.M., "Central Server Model for Multiprogrammed Computer Systems With Different Classes of Jobs," IBM Journal of Research and Development, 314-320 (1975).

[9] Cohen, J.W., "On Regenerative Processes in Queueing Theory," *Lecture Notes, Economics and Mathematics*, 121 (Springer-Verlag, Berlin, 1976).

[10] Cohen, J.W., "On Up- and Downcrossings," Journal of Applied Probability, *14*, 405-410 (1977).

[11] Cooper, R.B., *Introduction To Queueing Theory*, (second edition, North Holland, New York 1981). 1972).

[12] Cohen, J.W. and M. Rubinovitch, "On Level Crossings and Cycles in Dam Processes," Mathematics of Operations Research, *2*, 297-310 (1977).

[13] Crane, M.A. and A.J. Lemoine, *An Introduction to the Regenerative Method for Simulation Analyses* (Spring-Verlag, New York, 1977).

[14] Courtois, P.J. and J. Georges, "On a Single-Server Finite Queueing Model with State-Dependent Arrival and Service Processes," Operations Research, *19*, 424-435 (1971).
[15] Disney, R.L., D.C. McNickle, and B. Simon, "The M/G/1 Queue With Instantaneous Bernouli Feedback," Naval Research Logistics Quarterly, *27*, 635-644 (1980).
[16] Franken, P., "Some Applications of the Theory of Stochastic Point Processes in Queueing Theory, Mathematische Nachrichten, *70*, 309-316 (1976).
[17] Gross, D. and C.M. Harris, *Fundamentals of Queueing Theory*, (John Wiley and Sons, New York (1974).
[18] Heyman, D.P. and S. Stidham, "The Relation Between Customer and Time Averages in Queues," Operations Research, *28*, 983-994 (1980).
[19] Jain, J.L. and K. Seth, "A Unified Approach for Finding the Relationship Between the Continuous Time and the Embedded Markov Limiting Distributions for a Class of Queueing Problems," Paper presented at the Ninth Conference on Stochastic Processes and Their Applications, Evanston, Ill. (1979).
[20] Konig, D., T. Rolski, V. Schmidt, and D. Stoyan, "Stochastic Processes with Embedded Marked Point Processes and Their Application in Queueing," Mathematische Operationsforschung und Statistik Series Statistics, *9*, 125-141 (1978).
[21] Kuczura, A., "Piecewise Markov Processes," SIAM Journal of Applied Mathematics, *24*, 169-181 (1973).
[22] Neuts, M.F., "An Algorithmic Solution to the GI/M/c Queue With Group Arrivals," Cahiers du Centre de Recherche Opérationnelle, *21*, 109-119 (1979).
[23] Shanthikumar, J.G. "On Approximate Queueing Models of Dynamic Job Shops," Ph.D. Thesis, Department of Industrial Engineering, University of Toronto, Toronto, Canada (1979).
[24] Shanthikumar, J.G., "Some Analyses on the Control of Queues Using Level Crossings of Regenerative Processes," Journal of Applied Probability, *17*, 814-821 (1980).
[25] Shanthikumar, J.G., "Analysis of the Control of Queues with Shortest Processing Time Service Discipline," Journal of the Operations Research Society of Japan, *23*, 341-352 (1980).
[26] Shanthikumar, J.G., "Optimal Control of an M/G/1 Priority Queue Via *N*-Control, American Journal of Mathematical and Management Sciences, *1*, 191-212 (1981).
[27] Shanthikumar, J.G., "On the Buffer Behavior with Poisson Arrivals, Priority Service, and Random Server Interruptions," IEEE Transactions on Computers, *C30*, 781-786 (1981).
[28] Shanthikumar, J.G., "On Level Crossing Analysis of Queues," The Australian Journal of Statistics, *23*, 337-342 (1981).
[29] Shanthikumar, J.G., "Analysis of a Single Server Queue with Time- and Operation-Dependent Server Failures," Advances in Management Studies, *1* (1982).
[30] Shanthikumar, J.G. and R.G. Sargent, "An Algorithmic Solution for a Queueing Model of a Computer System with Interactive and Batch Jobs," Performance Evaluation, *1*, 344-357 (1981).
[31] Shanthikumar, J.G. and R.G. Sargent, "A Hybrid Simulation/Analytic Model of a Computerized Manufacturing System," Proceedings of the International Federation of Operational Research Societies Conference, 901-915 (1981).
[32] Shanthikumar, J.G. and C.C. Tien, "An Algorithmic Solution to Two-Stage Transfer Lines with Possible Scrapping of Units," Working Paper #81-006, Department of Industrial Engineering and Operations Research, Syracuse University, Syracuse, NY (1981).
[33] Takàcs, L., *Introduction to the Theory of Queues*, (Oxford University Press, New York 1962).

ON PATROLLING A CHANNEL

Alan R. Washburn

Naval Postgraduate School
Monterey, California

ABSTRACT

A simple formula is found to be just as accurate as a complicated one for estimating the probability of detection achievable by an ingenious searcher patrolling a channel or barrier. The difference between "detection" and "closure" is emphasized in an extension.

INTRODUCTION

A searcher whose speed is v and who can see any target within some distance R patrols a channel of width L looking for targets that come down the channel with speed u. By "patrols," we mean he selects any closed path and travels around it repeatedly; the endurance of the searcher is not an issue. The question is, "What is the best closed curve, and what fraction of the targets are detected?" If $v < u$, it is also of interest to find the fraction of targets that can be closed (touched) by the searcher, assuming that targets continue to move in a straight line down the channel. The problem is an old one, with the classic reference being [1]. In most naval applications, one role or the other is played by a submarine. The "channel" sometimes corresponds to that portion of a wide barrier that has been assigned to a particular searcher.

Our approach to this problem will be to first show that there is an upper bound on the probability of detection, corresponding to a particular infeasible patrol path for the searcher. We will then argue that there are feasible patrol paths for which the probability of detection is quite close to the upper bound. Finally, we shall provide a simple geometric analysis of the closure problem based on the infeasible patrol path.

AN UPPER BOUND ON THE DETECTION PROBABILITY

It is convenient to imagine that the targets are fixed to a tape moving down the channel with speed u, in which case the probability of detection for randomly distributed targets is just the ratio of the rate at which the searcher can examine new area on the tape to the rate uL at which new tape area appears. Let $w(t)$ be the speed of the searcher with respect to the tape at time t. Then the rate at which the searcher examines new area on the tape certainly does not exceed $2Rw(t)$. If the searcher is at the same position on his path at time T as at time 0, and if A is the amount of new area examined between 0 and T, then necessarily

$$A \leqslant 2R \int_0^T w(t)\,dt. \tag{1}$$

Let v $\cos\theta(t)$ be the component of the searcher's velocity in the same direction as the tape is moving. Since the searcher returns to his initial position at time T,

$$\int_0^T \cos\theta(t)\,dt = 0.$$

According to the Law of Cosines,

$$w(t) = \sqrt{u^2 + v^2 - 2u\,v\cos\theta(t)}.$$

The fact that

$$w(t) \leqslant \sqrt{u^2+v^2} - u\,\frac{v}{\sqrt{u^2+v^2}}\cos\theta(t)$$

is easy to show by observing that the right-hand side is positive and then squaring both sides. Consequently,

$$A \leqslant 2RT\sqrt{u^2+v^2}. \tag{2}$$

Since the amount of new tape that appears in time T is uLT, this leads to an upper bound on the probability of detection p:

$$p \leqslant \min\left\{1, (2R/L)\sqrt{1+(v/u)^2}\right\}. \tag{3}$$

CONVEYOR BELT SEARCH

In order to achieve the upper bound (3), it is evidently necessary that $\cos\theta(t) = 0$; that is, the searcher's motion must always be perpendicular to the sides of the channel. Unfortunately, the only way for the searcher to do this is to simply move back and forth across the channel, and this type of motion is inefficient for fast searchers because of the large amount of overlap with previously searched parts of the tape. The situation would be different if the tape were made into a tube by fastening the edges together, in which case the searcher could proceed perpendicularly to the target tracks without ever having to reverse his course. The effect of this would be the same as if there were a conveyor belt of searchers spaced L apart moving across the channel. Relative to such a belt, the tracks of the targets would differ from being normal to the belt by an angle ϕ whose tangent is v/u (see Figure 1). From Figure 1, it is apparent that a fraction $2R/L\cos\phi$ of all such tracks would encounter a searcher, assuming that tracks are equally likely to penetrate the belt anywhere, and also assuming $2R < L\cos\phi$. But this is the same as (3). Thus, a conveyor belt (CB) search of a channel produces the largest possible probability of detection. In most circumstances, however, a CB search is not feasible.

APPROXIMATELY OPTIMAL SEARCH PATHS

In order to avoid the overlap caused by the back and forth (BF) search, a fast searcher should move parallel to the channel near the edges. For example, he might choose a bow tie (BT) shaped path as in Figure 2. In [1], the choice was to make $D = R$ and α such that v $\sin\alpha = u$. The probability of detection was compared with that for $\alpha = 0$ (which corresponds to BF), and BT was found to be preferred to BF for fast searchers. Specifically, letting $r = v/u$ and $\lambda = L/2R - 1$, the results were:

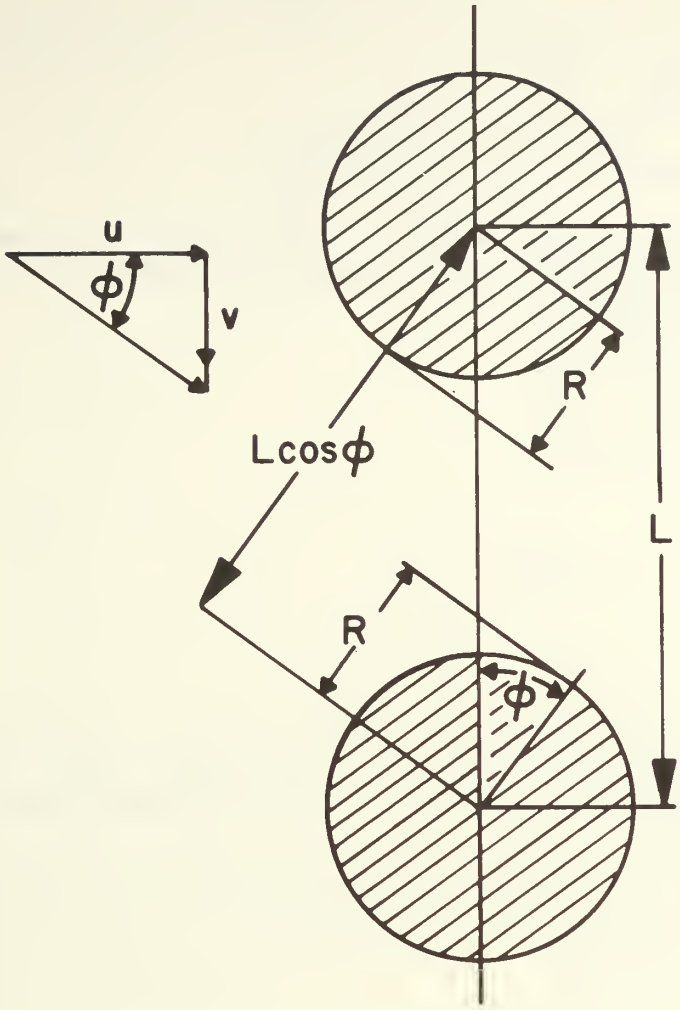


Figure 1 — Geometry of a Conveyor Belt Search

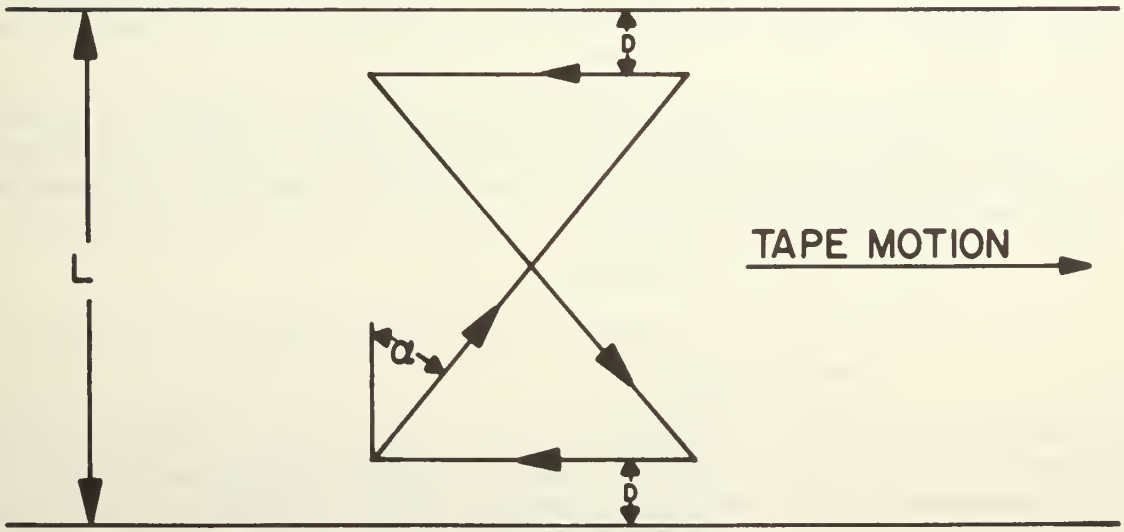


Figure 2 — The Bow Tie Search in a Channel

$$P_{BT} = \left(1 + \frac{r\sqrt{r^2 - 1}}{r + 1}\right) \frac{1}{\lambda + 1} \quad \text{for } r \geq 1 \tag{4}$$

$$P_{BF} = 1 - \left(\lambda - \frac{\sqrt{r^2 + 1} - 1}{2}\right)^2 \Big/ \lambda(\lambda + 1), \tag{5}$$

except that the probability cannot exceed 1 in either case. A proof of (4) and (5) is available in [3]. It is of interest to compare (4) with (3) when r is large and (5) with (3) when r is small. The analytic expression in (3) is

$$\frac{\sqrt{1 + r^2}}{\lambda + 1} \equiv P_{CB}$$

in the present notation. Letting

$$f_1 = \sqrt{1 + r^2}$$

and

$$f_2 = 1 + \frac{r\sqrt{r^2 - 1}}{r + 1},$$

it can be shown that

$$f_1^2 - f_2^2 = \frac{2r}{(1 + r)\ (r + \sqrt{r^2 - 1})}$$

from which it follows directly that $f_1 \geq f_2$ and $\lim_{r \to \infty}(2r^2)(f_1 - f_2) = 1$. Thus, $P_{CB} \geq P_{BT}$, and the difference becomes very small for large r. Also, it is simply a matter of expanding the square in (5) to show that

$$P_{BF} = P_{CB} - \left(\frac{\sqrt{r^2 + 1} - 1}{2}\right)^2 \Big/ \lambda(\lambda + 1).$$

Consequently, $P_{CB} \geq P_{BF}$, and

$$\lim_{r \to 0} \frac{4\lambda(\lambda + 1)}{r^2}(P_{CB} - P_{BF}) = 1;$$

i.e., the difference between the two becomes very small for small r. Figure 3 shows P_{CB},P_{BT}, and P_{BF} as a function of r for $\lambda = 2$.

The fact that $P_{CB} \geq \max(P_{BF}, P_{BT})$ does not follow directly from the fact that P_{CB} is an upper bound. The reason for this is that corners were "squared off" in [1] for ease of computation, which has the effect of computing too large a detection probability. The author has repeated these calculations with two changes:

(1) "round corners" were used

(2) all angles α such that $0 \leq v \sin \alpha \leq u$ were explored, and the optimal α selected for each (λ, r).

The computations are reported in detail in [4]. For $\lambda = 2$, the optimized detection probability is shown in Figure 3 as P_{OP}. The result is typical; P_{OP} is very nearly P_{CB} except when P_{CB} is large. Furthermore, it seems clear that even this shortfall could be eliminated by also permitting $D < R$ (see Figure 2), since the probability of detection for targets near the channel edges

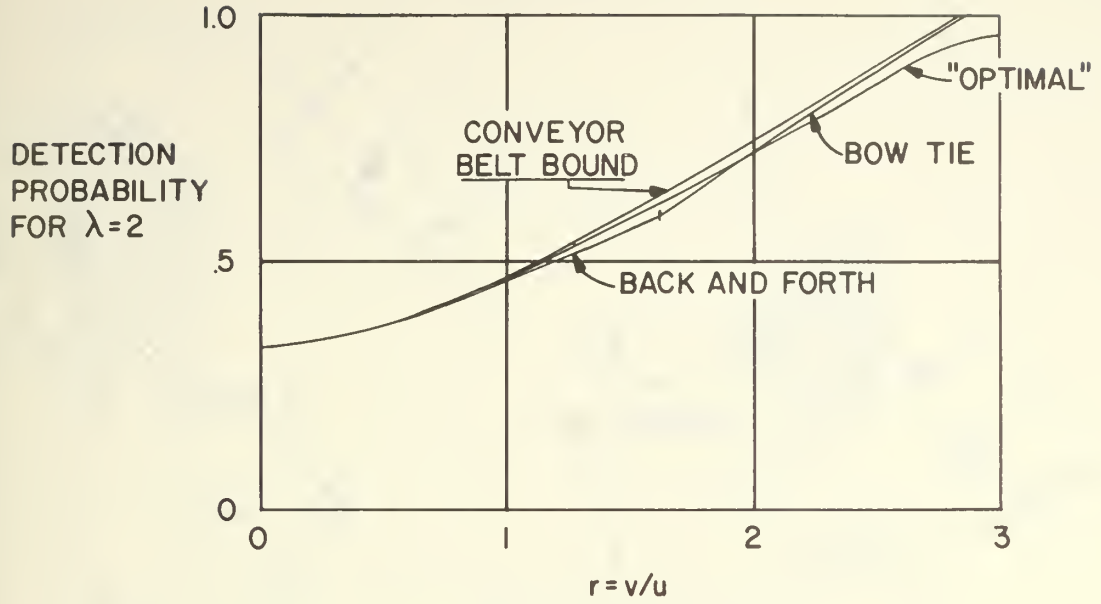


Figure 3 — Comparisons with the Conveyor Belt Bound

will be small as long as $D = R$. Erickson [2] has done some computations along these lines. Since the searcher is actually allowed to use any closed path whatever, rather than only the class of generalized bow ties that we have permitted him, it seems reasonable to come to the conclusion that there exists a search path that is close to the conveyor belt in terms of probability of detection, provided the probability of detection is not too close to 1. It follows from this that formula (3) is generally a better assumption than either (4) or (5), for reasons of both simplicity and accuracy. The geometric simplicity of the conveyor belt search also makes various extensions relatively easy. An example of such an extension is given in the next section.

THE PROBABILITY OF CLOSURE

Suppose there is a conveyor belt of searchers looking for targets that move in straight lines normal to the conveyor belt. Upon detecting a target, a searcher attempts to catch (close) it at some speed v′ that may exceed his patrol speed v. If $v' > u$, he will certainly succeed, but if $v' \leqslant u$, he will succeed only if the detection is made in a certain sector of the detection circle as illustrated in Figure 4, where $\cos \psi = v'/u$. A target will ultimately be caught if and only if it touches the shaded part of some detection circle. As in Figure 1, the targets arrive relative to the conveyor belt at angle ϕ. The geometry of computing the closure probability depends on the relative magnitudes of ϕ and ψ. Figure 4 illustrates the case where $\phi \leqslant \psi$, and Figure 5 the case where $\phi \geqslant \psi$. Noting that $\cos(\psi - \phi) + \cos(\psi + \phi) = 2 \cos \phi \cos \psi$, we see that

$$P_C = \begin{cases} (2R/L) \tan \phi & \text{if } v > u \\ (2R/L) \cos \psi & \text{if } v \leqslant u \text{ and } \phi \leqslant \psi \\ \dfrac{R}{L}\left\{\dfrac{1 + \cos(\psi + \phi)}{\cos \phi}\right\} & \text{if } v \leqslant u \text{ and } \phi \geqslant \psi. \end{cases} \tag{6}$$

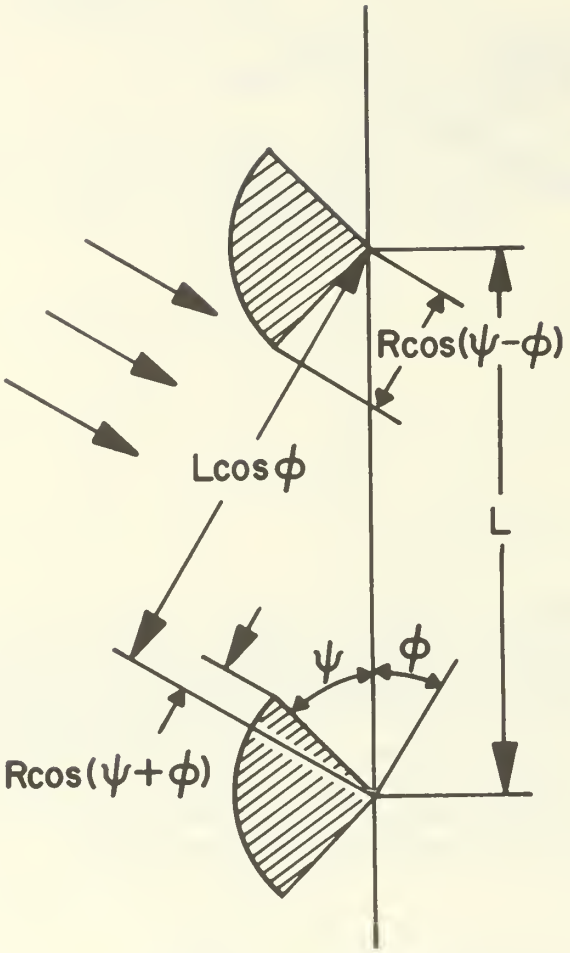


Figure 4 — Closure Geometry when $\phi \leqslant \psi$

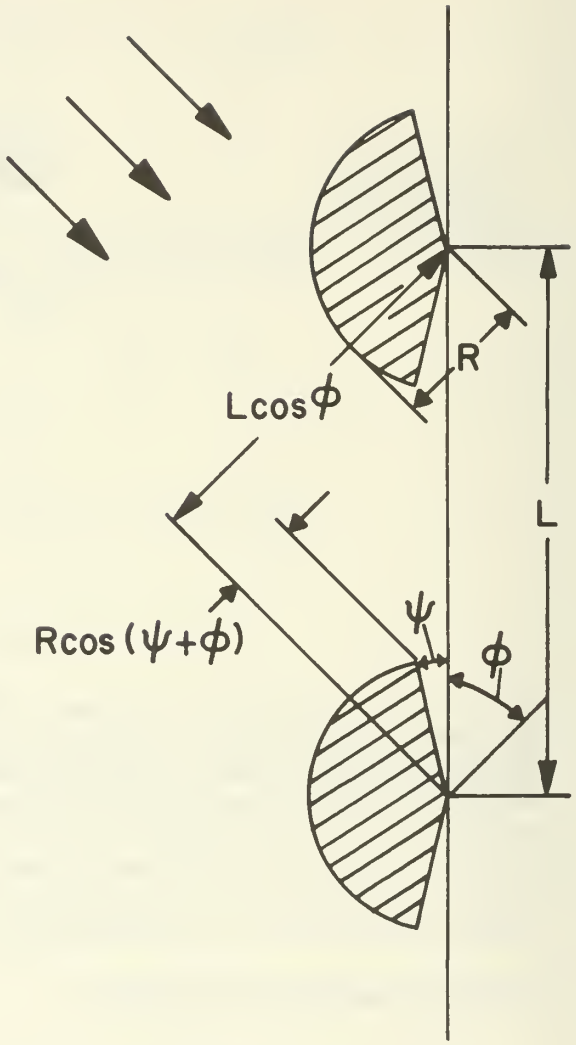


Figure 5 — Closure Geometry when $\phi \geqslant \psi$

Figure 6 shows P_C vs u/v when $v' = v$ and $R/L = .1$. The downward jump when $u/v = 1$ shows that there is a definite benefit to the target for being faster than the searcher. Of course, increases in u may also affect R, particularly if the detection mechanism is based on sound waves, so it does not follow that the best strategy for the target is to go as fast as possible.

REFERENCES

[1] Center for Naval Analyses, Arlington, VA, "Search and Screening," Operations Evaluation Group Report 56 (1946).

[2] Erickson, S.A., "Detection Rates for a Multi-Searcher Barrier and a Single Transistor," presented at Joint ORSA/TIMS Meeting, Las Vegas, NV (1975).

[3] Hetzler, J.S., Jr., "The Application of Operations Analysis to Weapons Systems Development," NOLTR 69-154, U.S. Naval Ordnance Laboratory, White Oak, MD (1969).

[4] Washburn, A.R., "Patrolling a Channel Revisited," NPS55WS 75121 Report, Naval Postgraduate School, Monterey, CA (1976).

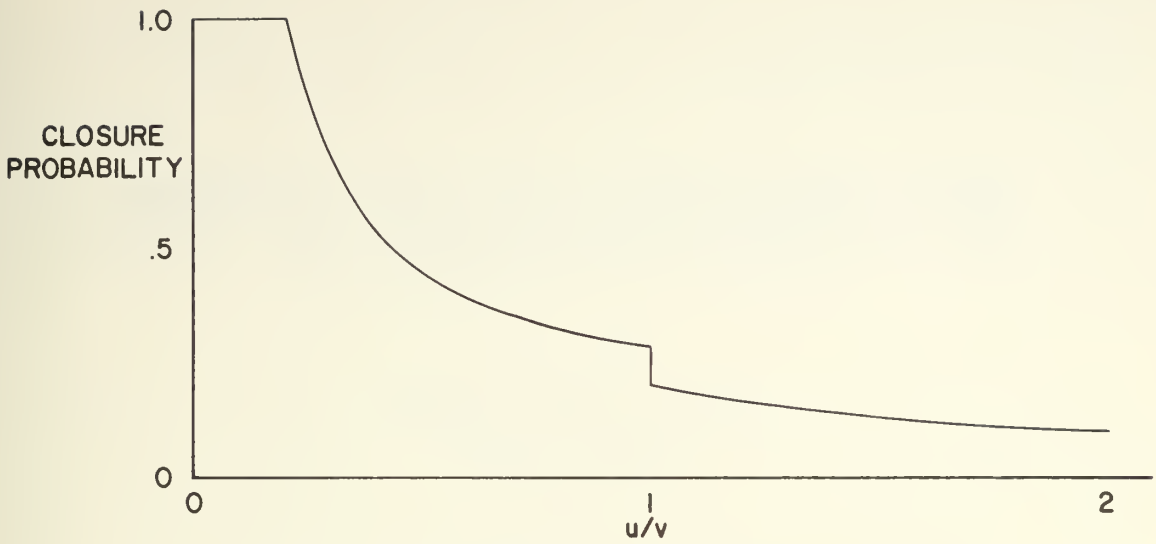


Figure 6 — Dependence of Closure Probability on Target Speed

THEORETICAL ANALYSIS OF THE GENERAL LINEAR MODEL FOR LANCHESTER-TYPE COMBAT BETWEEN TWO HOMOGENEOUS FORCES*

James G. Taylor

Department of Operations Research
Naval Postgraduate School
Monterey, California

ABSTRACT

This paper studies Lanchester-type combat between two homogeneous forces modeled by the so-called general linear model with continuous replacements/withdrawals. It demonstrates that this model can be transformed into a simpler canonical form, which is also shown to arise from fixed-force-level-breakpoint battles modeled by Lanchester-type equations for modern warfare. Analytical expressions for the force levels for the general variable-coefficient linear model with continuous replacements/withdrawals are constructed out of so-called general Lanchester functions for the model without replacements/withdrawals, for which all solutions are shown to be nonoscillatory in the strict sense. These force-level results are unfortunately so complicated and opaque that the constant-coefficient version of the model must be studied before any insights into the dynamics of combat may be analytically obtained. Thus, fairly complete results are given for the general linear model with constant attrition-rate coefficients and constant rates of replacement/withdrawal. However, the expressions for the force levels are still so complicated that we have not been able to develop battle-outcome prediction conditions directly from them alone but have had to establish general results on the qualitative behavior of solutions. A significant result (and one that greatly complicates the prediction of battle outcome) is that all solutions to the model with replacements/withdrawals are no longer necessarily nonoscillatory in the strict sense, i.e., both sides' force levels can take on negative values if the force-on-force attrition equations are not "turned off" at the right time. Thus, this paper shows that the addition of continuous replacements/withdrawals to a Lanchester-type model may significantly change the qualitative behavior of the force-level trajectories. Battle-outcome prediction conditions are nevertheless given, and important insights into the dynamics of combat are briefly indicated.

1. INTRODUCTION

Deterministic Lanchester-type combat models (so-called after pioneering work by F. W. Lanchester [12]) are widely used today in both the United States [16], [19], [22] and also

*This research was partially supported by the Headquarters U.S. Air Force, Studies and Analysis Group (HQUSAF/SAZ) and partially by the Office of Naval Research as part of the Foundation Research Program at the Naval Postgraduate School.

NATO countries [8], [9] for defense-planning purposes. The modern high-speed, large-scale computer has made possible the development and use [2] of quite complex differential equation-based operational models of combat operations (e.g., see [4]). It is indeed surprising, therefore, that relatively little effort has been devoted to the mathematical analysis of various Lanchester-type paradigms (see [15, p. 88]) from which these complex computer-based combat models have been developed by the process of model enrichment (e.g., see [3], [13], or [16]). Accordingly, this paper will theoretically investigate the qualitative behavior of a relatively simple Lanchester-type model with replacements to develop some important insights into the dynamics of combat.

How do replacements and withdrawals affect the course of a battle? How fast should troops be put ashore in an amphibious landing to insure its success? Will reinforcements reach NATO in time? These are important questions both for the military tactician and also the defense planner. In this paper we will study a relatively simple Lanchester-type model that explicitly considers continuous replacements and/or withdrawals in order to provide some basic insights into such issues as delineated above. Our basic analysis strategy is to use a *simplified auxiliary model* for understanding the basic dynamics and behavior of a large-scale *complex operational model* [6], [16]. This has been the basic motivation for studying the relatively simple Lanchester-type paradigm considered here. Also, many times one can fit such a simple analytical model to data generated by a detailed combat simulation, and thus a simple analytical model such as we consider here may provide an economical framework for summarizing simulation output data [5], [10].

Thus, this paper theoretically investigates a relatively simple Lanchester-type model that explicitly considers replacements and/or withdrawals. Essentially, no analytical work (except for the rather dated work of Morse and Kimball [14] and Karns [11]) has been previously done on such models. In particular, no previous theoretical work has been reported on the qualitative behavior of solutions to such models. Our new theoretical results show that the qualitative behavior of force-level trajectories is fundamentally changed by the inclusion of continuous replacements and/or withdrawals in a Lanchester-type combat model.

2. THE GENERAL LINEAR MODEL FOR LANCHESTER-TYPE COMBAT BETWEEN TWO HOMOGENEOUS FORCES

In this section we will briefly examine the general linear-differential equation model for combat between two homogeneous forces. Special cases of this general model will be examined in more detail in subsequent sections of this paper.

Accordingly, we consider the following Lanchester-type equations for x and $y > 0$

$$\text{(2.1)} \qquad \begin{cases} \dfrac{dx}{dt} = -a(t)y - \beta(t)x + r(t) \text{ with } x(0) = x_0, \\ \dfrac{dy}{dt} = -b(t)x - \alpha(t)y + s(t) \text{ with } y(0) = y_0, \end{cases}$$

where $x(t)$ and $y(t)$ denote the X and Y force levels at time t, and $a(t)$ and $b(t)$ denote Lanchester attrition-rate coefficients, which represent the fire effectiveness of a single "typical" firer on each side. There are several different sets of physical and operational circumstances to which the model (2.1) may be hypothesized to apply, and the coefficients $\alpha(t)$, $\beta(t)$, $r(t)$, and $s(t)$ have different physical/operational interpretations, depending upon the context in which the model is viewed. We will now discuss several such possibilities.

The term $r(t)$ in the first of Equations (2.1) can model either (I) the replacement rate for the X force (with a negative value representing a net continuous withdrawal), or (II) the attrition [in which case $r(t)<0$] of the X force from exogenous fires (not subject to attrition) at a rate not dependent on X's force level. Similar remarks apply to $s(t)$. For simplicity, however, we will consider only the first possibility here and will consequently refer to $r(t)$ and $s(t)$ as replacement rates. Even within this context, two different tactical situations may again be hypothesized to yield the above equations (2.1) (cf. Figure 4 of Taylor [19]):

either (S1) "aimed-fire" combat between two homogeneous forces with "operational" losses and with continuous replacements,

or (S2) "aimed-fire" combat between two homogeneous primary forces (or infantries) with superimposed effects of supporting fires not subject to attrition and with continuous replacements for the primary forces (see Figure 1).

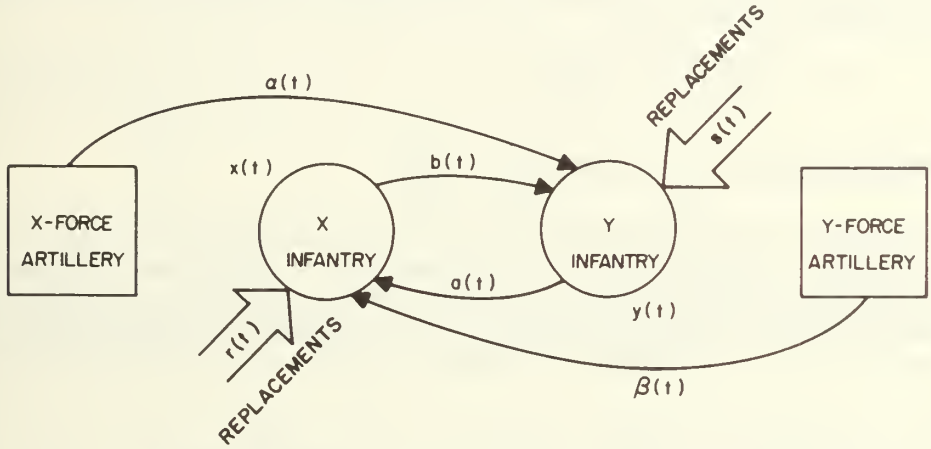


FIGURE 1. "Aimed-fire" combat between two homogeneous primary forces (infantries) with superimposed effects of supporting fires (here, from artillery) not subject to attrition and with continuous replacements for the primary forces.

In the second case (S2), it is assumed that each side uses "aimed" fire and that target-acquisition times do not depend on the number of enemy targets. The supporting weapons are assumed to employ "area" fire against enemy infantry (see Taylor [19] for a more thorough discussion of modeling assumptions; also [17], [20]). In this case, determination of numerical values for the attrition-rate coefficients $\alpha(t)$ and $\beta(t)$, modeling the effects of the supporting fires, follows along the lines discussed in Taylor [19, Section 6.3] . In the simplest instance we then have that, for example, $\alpha(t) = a_{L_U} \nu_U u_0 / A_Y$, where the X force's artillery is denoted as U, a_{L_U} denotes the lethal area of a single U artillery round, ν_U denotes the U firing rate per tube, u_0 denotes the U force level (which is constant because it is assumed that the U force suffers no losses), and A_Y denotes the presented area of the region occupied by the Y force.

Mathematically, we make the following assumptions about the attrition-rate coefficients and replacement rates in the model (2.1):

(A1) $a(t)$ and $b(t)$ are defined, positive, and continuous for $t \in (t_0, +\infty)$ with $t_0 \leqslant 0$,

(A2) $\alpha(t)$ and $\beta(t) \geqslant 0$ for $t \in [t_0, +\infty)$,

(A3) $a(t)$, $b(t)$, $\alpha(t)$, $\beta(t)$, $r(t)$, and $s(t) \in L(t_0, T)$ for any finite $T > t_0$.

Here the notation $a(t) \in L(t_0, T)$ is used to mean that $\int_{t_0}^{T} a(t)\,dt$ exists (and is given by a finite quantity). Further details are to be found in [17], [20]. It follows from (A3) that any force-level trajectory to (2.1) for a given set of initial conditions is continuous and unique. We will place no further restrictions on the replacement rates $r(t)$ and $s(t)$, and, consequently, negative values are possible for them. We will further assume, however, that $a(t)$ and $b(t)$ are given in the form [18], [20]

$$a(t) = k_a g(t) \text{ and } b(t) = k_b h(t), \tag{2.2}$$

where k_a and k_b are positive constants chosen so that $a(t)/b(t) \equiv k_a/k_b$ when $\partial\{a(t)/b(t)\}/\partial t \equiv 0$. In other words, k_a and k_b are basically "scale factors," which are useful for parametric study of battle outcome as related to various system parameters [18] (see also [20]). It is then convenient to introduce the *combat-intensity parameter* λ_I and the *relative-fire-effectiveness parameter* λ_R defined by

$$\lambda_I = \sqrt{k_a k_b} \text{ and } \lambda_R = \frac{k_a}{k_b}. \tag{2.3}$$

3. TRANSFORMATION TO A CANONICAL FORM

No results have previously appeared in the literature for the general model (2.1) with variable attrition-rate coefficients. We will now show that (2.1) may be transformed into a simpler canonical form to which results [16], [18] for variable-coefficient Lanchester-type equations of modern warfare (3.5) may be applied. Thus, this model (3.5) is really basic for studying a wide variety of combat situations (cf. Sections 5.4 and 5.5 of Taylor [16]). The substitution

$$p(t) = x(t) \exp\left\{\int_0^t \beta(s)\,ds\right\}, \quad q(t) = y(t) \exp\left\{\int_0^t \alpha(s)\,ds\right\}, \tag{3.1}$$

transform (2.1) into

$$\begin{cases} \dfrac{dp}{dt} = -A(t)q + V(t) \text{ with } p(0) = x_0, \\ \dfrac{dq}{dt} = -B(t)p + W(t) \text{ with } q(0) = y_0, \end{cases} \tag{3.2}$$

where

$$A(t) = a(t) \exp\left\{\int_0^t [\beta(s) - \alpha(s)]\,ds\right\},$$

$$B(t) = b(t) \exp\left\{-\int_0^t [\beta(s) - \alpha(s)]\,ds\right\}, \tag{3.3}$$

and

$$V(t) = r(t) \exp\left\{\int_0^t \beta(s)\,ds\right\}, \quad W(t) = s(t) \exp\left\{\int_0^t \alpha(s)\,ds\right\}. \tag{3.4}$$

The transformation (3.1) is motivated by looking for an "integrating factor" for, for example, the first equation of (2.1), as writing $dx/dt + \beta(t)x = -a(t)y + r(t)$ suggests to one.

As we have seen above in Section 2, we may consider Equations (3.2) to model "aimed fire" combat between two homogeneous forces with continuous replacements. However, there is another very important set of circumstances that leads to equations of this form. Consider aimed-fire combat between two homogeneous forces modeled by Lanchester's equations for modern warfare (for x and $y > 0$)

$$\begin{cases} \dfrac{dx}{dt} = -a(t)y \text{ with } x(0) = x_0, \\ \dfrac{dy}{dt} = -b(t)x \text{ with } y(0) = y_0. \end{cases} \tag{3.5}$$

In this model the state variables $x(t)$ and $y(t)$ are the numbers of combatants that are effective on each side. Consider now a fixed-force-level-breakpoint battle [16], [19] . If we introduce new state variables $X(t)$ and $Y(t)$ defined by

$$X(t) = x(t) - x_{BP} \text{ and } Y(t) = y(t) - y_{BP}, \tag{3.6}$$

where x_{BP} and y_{BP} denote the X and Y force-level breakpoints, then

$$\begin{cases} \dfrac{dX}{dt} = -a(t)Y - \mathrm{v}(t) \text{ with } X(0) = x_0 - x_{BP}, \\ \dfrac{dY}{dt} = -b(t)X - w(t) \text{ with } Y(0) = y_0 - y_{BP}, \end{cases} \tag{3.7}$$

where $\mathrm{v}(t) = a(t)y_{BP}$ and $w(t) = b(t)x_{BP}$. These equations (3.7) have the same form as (3.2), and thus we see that the equations (3.2) may also be taken to model force attrition "above a unit's breakpoint." We observe that for the transformed force-level variable X defined by (3.6), $X = 0$ corresponds to the X force reaching its breakpoint.

The force-level trajectories $x(t)$ and $y(t)$ for the model (2.1) [equivalently, (3.2) or (3.7)], however, no longer possess a very important mathematical property that is possessed [18] by all solutions to (3.5) with $a(t)$ and $b(t) \geqslant 0$ for all $t \geqslant 0$ and x_0 and $y_0 > 0$: namely, all solutions to (2.1) are no longer nonoscillatory in the strict sense that $x(t)$ and $y(t)$ can how have more than one zero between them. Here it is convenient [and indeed usually implicitly done when one works with analytical solutions such as (6.1.1) and (6.1.2) below] to relax the nonnegativity restrictions on the force levels. The possible lack of such nonoscillatory behavior will be shown by considering a specific example below. This mathematical property is troublesome and makes analysis of battles modeled with (2.1) much more difficult than those modeled with (3.5).

4. REPRESENTATION OF FORCE LEVELS

The solution to the inhomogeneous differential-equation combat model (2.1) may be constructed out of the same "building blocks" that may be used to construct the solution to the corresponding homogeneous system of linear differential equations. Thus, the X force level as a function of time, $x(t)$, for the general model (2.1) may be represented as

$$x(t) = \left[\exp\left\{-\int_0^t \beta(s)\,ds\right\}\right]\left[x_0\left\{C_Q(0)C_P(t) - S_Q(0)S_P(t) - y_0\sqrt{\lambda_R}\{C_P(0)S_P(t) - S_P(0)C_P(t)\right\}\right.$$

$$\left. + \sqrt{\lambda_R}\int_0^t \frac{Z(s)}{z(s)}\{C_P(s)S_P(t) - S_P(s)C_P(t)\}\,ds\right], \tag{4.1}$$

where $Z(t) = -A(t)S(t) + dR/dt - \{R(t)/A(t)\}dA/dt = \{-a(t)s(t) + dr/dt - r(t)[(1/a(t))da/dt - \alpha(t)\}\exp\int_0^t \beta(s)\,ds\}$, and the hyperbolic-like general Lanchester functions (GLF) $C_P(t)$, $S_P(t)$, $C_Q(t)$, and $S_Q(t)$ arise from the corresponding homogeneous system of linear differential equations, which is considered in the next section. Thus, they hyperbolic-like GLF $C_P(t)$ and $S_P(t)$ are linearly-independent solutions to the P force-level equation (5.4) that satisfy the initial conditions (5.5), and $C_Q(t)$ and $S_Q(t)$ are similarly defined. The above result (4.1) is readily developed by considering (3.1) and applying well-known results for inhomogeneous linear ordinary differential equations (e.g., see Hildebrand [7, pp. 29-30]) combined with those for the model (5.1) [20].

5. THE MODEL WITHOUT REPLACEMENTS AND WITHDRAWALS (COMBAT WITH SUPPORTING FIRES)

An important special case of the general linear combat model (2.1) is that in which there are no replacements and withdrawals, i.e., $r(t)$ and $s(t) \equiv 0$, and in this case our combat model becomes

$$\begin{cases} \dfrac{dx}{dt} = -a(t)y - \beta(t)x \text{ with } x(0) = x_0, \\ \dfrac{dy}{dt} = -b(t)x - \alpha(t)y \text{ with } y(0) = y_0. \end{cases} \tag{5.1}$$

An essential characteristic of all solutions to (5.1) (if we forget for the time being about the nonnegativity restrictions on x and y) is their so-called nonoscillation (here in the strict sense).

THEOREM 1: At most, one of the two force levels $x(t)$ and $y(t)$ for the model (5.1) can ever vanish in finite time.

PROOF: Multiplying the first of Equations (5.1) by y, the second by x, and adding; we obtain after some straightforward algebraic manipulations

$$\frac{d}{dt}(xy) + \{\alpha(t) + \beta(t) + 2\sqrt{a(t)b(t)}\}xy = -\{\sqrt{a(t)}y - \sqrt{b(t)}x\}^2.$$

whence integration yields

$$x(t)y(t) = \left\{\exp\left[-\int_0^t \{\alpha(s) + \beta(s) + 2\sqrt{a(s)b(s)}\}\,ds\right]\right\}\left\{x_0y_0\right.$$

$$\left. - \int_0^t \left\{\sqrt{a(s)}y(s) - \sqrt{b(s)}x(s)\right\}^2 \exp\left[\int_0^s \{\alpha(r) + \beta(r) + 2\sqrt{a(r)b(r)}\}\,dr\right]ds\right\}. \tag{5.2}$$

The theorem now follows by observing that the second multiplicative factor on the right-hand side of (5.2) determines the sign of $x(t)y(t)$ and that this factor can change sign at most once. The simultaneous vanishing of both $x(t)$ and $y(t)$ is precluded by the fact that if $x(t_a) = y(t_a) = 0$ for some finite $t_a > 0$, then the uniqueness of solutions to (5.1) would imply that $x(t) \equiv y(t) \equiv 0$.

Q.E.D

The significance of Theorem 1 is that the mathematical behavior of the force levels is fundamentally different for the models (2.1) and (5.1): we show in Section 6.2 below by the example of the constant-coefficient version of (2.1) that (again ignoring the nonnegativity restriction on the force levels) both $x(t)$ and $y(t)$ can become zero (and then negative) for the general linear model (2.1). Thus, the addition of replacements and/or withdrawals to the model (5.1), i.e., take $r(t)$ or $s(t) \not\equiv 0$ in (2.1), leads to fundamentally different behavior for the force-level trajectories.

Another reason for briefly considering this model without replacements and withdrawals (5.1) is that the same basic building blocks may be used to construct force-level solutions to both (2.1) and (5.1). Thus, the X force level as a function of time, $x(t)$, for the model (5.1) may be written as [20]

$$x(t) = \left[\exp\left\{-\int_0^t \beta(s)\,ds\right\}\right]\left[x_0\{C_Q(0)\,C_P(t) - S_Q(0)\,S_P(t)\}\right.$$

$$\left. -y_0\sqrt{\lambda_R}\{C_P(0)\,S_P(t) - S_P(0)\,C_P(t)\}\right], \tag{5.3}$$

where the hyperbolic-like GLF $C_P(t)$ and $S_P(t)$ are linearly-independent solutions to the *P force-level equation*

$$\frac{d^2p}{dt^2} - \left\{\beta(t) - \alpha(t) + \frac{1}{a(t)}\,\frac{da}{dt}\right\}\frac{dp}{dt} - a(t)\,b(t)\,p = 0, \tag{5.4}$$

with initial conditions

$$C_P(t_0) = 1, \qquad S_P(t_0) = 0,$$
$$\frac{1}{a(t_0)}\,\frac{dC_P}{dt}(t_0) = 0, \qquad \frac{1}{a(t_0)}\,\frac{dS_P}{dt}(t_0) = \frac{1}{\sqrt{\lambda_R}}. \tag{5.5}$$

The GLF $C_Q(t)$ and $S_Q(t)$ are similarly defined. It should be noted here that the above result (5.3) is a special case of (4.1). Further details and analysis of the model without replacements and withdrawals (5.1) may be found in [17] and [20].

6. THE GENERAL LINEAR MODEL WITH CONSTANT ATTRITION-RATE COEFFICIENTS AND REPLACEMENTS

In the case of constant attrition-rate coefficients and replacements, the general linear model (2.1) reads

$$\begin{cases} \dfrac{dx}{dt} = -ay - \beta x + r \text{ with } x(0) = x_0, \\ \dfrac{dy}{dt} = -bx - \alpha y + s \text{ with } y(0) = y_0, \end{cases} \tag{6.1}$$

where a, b, α, β, r, and s denote quantities that remain constant during a particular battle, and we assume that a and $b > 0$, while α and $\beta \geqslant 0$. Although there are several different sets of physical circumstances that may be hypothesized to yield (6.1) (see Section 2 above), we will consider (6.1) to model "aimed-fire" combat between two homogeneous forces with supporting fires not subject to attrition and continuous replacements/withdrawals. In this case we should consider r and s to be replacement rates, with a negative value denoting a net rate of withdrawal of forces. Accordingly, we will place no restrictions on the replacement rates r and s, i.e., r and s are unrestricted in sign.

The model (6.1) is of interest because it provides insights into the consequences of additional troops (continuously) committed to battle. We may consider a term like, for example, r to represent the rate at which additional X forces are committed to battle. Another related interpretation is that r represents the net rate at which the X force enters the fields of fire of the Y force. Such interpretations essentially apply to small-unit combat in the fire fights. We may also (see Section 2 above) however, consider (6.1) to model *combat with operational losses and continuous replacements.* In this case we may consider (6.1) to apply to *large-scale combat* over a sustained period of time, and then r and s represent the rates at which additional resources are committed to the theater of operations (see Morse and Kimball [14, pp. 71-73]). In this light, analysis of this combat model will provide important insights into the nature of tradeoffs among (1) direct combat capability, (2) "build-up" capability, and (3) operational losses. In terms of the NATO scenario, the model (6.1) provides rough insights into the structure of tradeoffs among the quality of weapon systems, the quantity of weapon systems, and the "build-up" rates at which new systems are introduced into the theater of operations.

6.1. Force-Level Results

Unlike the previous variable-coefficient versions considered above, the constant-coefficient model (6.1) yields an analytical solution that is simple enough to provide some important insights into the dynamics of combat through direct analysis. When $ab \neq \alpha\beta$, the X and Y force levels $x(t)$ and $y(t)$ for the model (6.15.1) are given by*

$$x(t) = \xi + Ae^{(\theta-\sigma)t} + \left(\frac{\theta+\delta}{b}\right)Be^{-(\theta+\sigma)t}, \tag{6.1.1}$$

and

$$y(t) = \eta - \left(\frac{\theta+\delta}{a}\right)Ae^{(\theta-\sigma)t} + Be^{-(\theta+\sigma)t}, \tag{6.1.2}$$

where

$$A = \frac{ab}{2\theta(\theta+\delta)}\left\{(x_0-\xi) - \left(\frac{\theta+\delta}{b}\right)(y_0-\eta)\right\}, \tag{6.1.3}$$

*An equivalent analytical result has been given by Morse and Kimball [14, p. 72]. However, their result is not in a particularly convenient form for determining the qualitative behavior of the model (6.1). Moreover, the qualitative behavior itself of such solutions (as delineated in Section 6.2 below) was not known at that time, and consequently incorrect battle-outcome prediction conditions were implied in [14, p. 72]. Furthermore, the solution behavior shown by the example in Table 1 was not detected by Morse and Kimball.

$$B = \frac{ab}{2\theta(\theta+\delta)}\left\{\left[\frac{\theta+\delta}{a}\right](x_0-\xi)+(y_0-\eta)\right\}, \tag{6.1.4}$$

$$\xi = \frac{as-\alpha r}{\Delta},\ \eta = \frac{br-\beta s}{\Delta},\ \Delta = ab-\alpha\beta, \tag{6.1.5}$$

$$\theta = \sqrt{ab+\delta^2},\ \delta = \frac{\beta-\alpha}{2},\ \text{and}\ \sigma = \frac{\alpha+\beta}{2}. \tag{6.1.6}$$

Let us also note the following identity

$$\frac{\theta+\delta}{b} = \sqrt{R}\left\{\frac{S}{2}+\sqrt{\left[\frac{S}{2}\right]^2+1}\right\} \tag{6.1.7}$$

where

$$R = \frac{a}{b}\ \text{and}\ S = \frac{\beta-\alpha}{\sqrt{ab}}. \tag{6.1.8}$$

The parameters R and S have interesting military interpretations, and the interested reader can find a discussion about them in Taylor and Parry [21, p. 527], Taylor [16, p. 789], or Taylor [17, p. 368].

When $ab = \alpha\beta$, the X and Y force levels $x(t)$ and $y(t)$ for the model (6.1) are given by

$$x(t) = x_0e^{-(\alpha+\beta)t} + \left[\frac{\alpha r-as}{\alpha+\beta}\right]t + \left\{\frac{(\beta r+as)}{(\alpha+\beta)^2}+\left[\frac{\alpha x_0-ay_0}{\alpha+\beta}\right]\right\}\{1-e^{-(\alpha+\beta)t}\}, \tag{6.1.9}$$

and

$$y(t) = y_0e^{-(\alpha+\beta)t} - \left[\frac{b}{\alpha}\right]\left[\frac{\alpha r-as}{\alpha+\beta}\right]t + \left\{\frac{(\alpha s+br)}{(\alpha+\beta)^2}-\left[\frac{b}{\alpha}\right]\left[\frac{\alpha x_0-ay_0}{\alpha+\beta}\right]\right\}\{1-e^{-(\alpha+\beta)t}\}. \tag{6.1.10}$$

In this latter case, i.e., when $ab = \alpha\beta$, the constant-coefficient combat model (6.1) possesses the state equation

$$b(x_0-x) = \beta(y_0-y)+(\beta s-br)t, \tag{6.1.11}$$

which yields the overall casualty-exchange ratio is constant, i.e.,

$$\frac{x_c}{y_c} = \frac{\beta}{b}, \tag{6.1.12}$$

where the X and Y casualties are given by

$$x_c = x_0+rt-x,\ \text{and}\ y_c = y_0+st-y. \tag{6.1.13}$$

Let us observe that in all cases the instantaneous casualty-exchange ratio dx/dy is given by

$$\frac{dx}{dy} = \frac{\beta}{b} + \left\{ \frac{r - \frac{\beta}{b} s - \frac{\Delta}{b} y}{s - bx - \alpha y} \right\}, \tag{6.1.14}$$

which for $ab = \alpha\beta$ becomes

$$\frac{dx}{dy} = \frac{\beta}{b} + \left\{ \frac{r - (\beta/b)s}{s - bx - \alpha y} \right\}. \tag{6.1.15}$$

In particular, for $br = \beta s$ and $ab = \alpha\beta$ we have the linear law

$$b(x_0 - x) = \beta(y_0 - y). \tag{6.1.16}$$

Let us finally sketch the development of the above expressions for the force levels $x(t)$ and $y(t)$. When $ab \neq \alpha\beta$, we may write (6.1) as

$$\begin{cases} \frac{dx}{dt} = -a(y - \eta) - \beta(x - \xi) \\ \frac{dy}{dt} = -b(x - \xi) - \alpha(y - \eta), \end{cases} \tag{6.1.17}$$

whence the substitution

$$X(t) = x(t) - \xi \text{ and } Y(t) = y(t) - \eta \tag{6.1.18}$$

transforms (6.1.17) into

$$\begin{cases} \frac{dX}{dt} = -aY - \beta X \text{ with } X(0) = x_0 - \xi, \\ \frac{dY}{dt} = -bX - \alpha Y \text{ with } Y(0) = y_0 - \eta, \end{cases} \tag{6.1.19}$$

for which we have given a solution, i.e., expressions for $X(t)$ and $Y(t)$, in [20, p. 118] (see also Bach, Dolansky, and Stubbs [1]). When $ab = \alpha\beta$, we may write (6.1) as

$$\begin{cases} \frac{dx}{dt} = r - \beta\left[x + \frac{\alpha}{b} y\right] \text{ with } x(0) = x_0, \\ \frac{dy}{dt} = s - b\left[x + \frac{\alpha}{b} y\right] \text{ with } y(0) = y_0, \end{cases} \tag{6.1.20}$$

whence follows (6.1.11). The state equation (6.1.11) may then be used, for example, to eliminate y from the first of Equations (6.1.20), and hence the expression (6.1.9) for $x(t)$ may be obtained by a simple integration.

6.2. Theoretical Analysis of Solution Behavior

Determination of the qualitative behavior, e.g., battle-outcome prediction conditions, for the linear combat model with replacements (6.1) is much more difficult than we have heretofore encountered because the force levels $x(t)$ and $y(t)$ no longer possess a very important

mathematical property that facilitated analysis of combat modeled with Lanchester's equations for modern warfare (3.5): namely, all solutions to (6.1) are no longer nonoscillatory in the strict sense that $x(t)$ and $y(t)$ can have more than one zero. We will give an example of such solution behavior below. However, analysis of the qualitative behavior of the model (6.1) is relatively straightforward when $ab > \alpha\beta$, i.e., the intensity of combat between the primary systems exceeds the "intensity" of the supporting fires, although the above analytical results for $x(t)$ and $y(t)$, e.g., (6.1.1) and (6.1.2), are nevertheless complicated enough so that direct analysis of them has not proven completely fruitful. We will now develop *battle-outcome prediction conditions* for a fight-to-the-finish for this case in which $ab > \alpha\beta$.

We begin with a few heuristics to provide some intuitive understanding of our results. Let us first observe that $\theta - \sigma > 0$ if and only if $ab > \alpha\beta$. Hence, in this case the exponential $e^{(\theta-\sigma)t}$ in (6.1.1) and (6.1.2) is a strictly increasing function that grows without bound. Furthermore, the signs of $x(t)$ and $y(t)$ for large t are opposite and determined by the sign of A. However, further analysis and a few additional assumptions are required to convert these preliminary heuristics into mathematically precise battle-outcome prediction results. It is worthwhile to finally observe that then $A = 0$, i.e., $(x_0 - \xi) = (y_0 - \eta)(\theta + \delta)/b$, (6.1.1) and (6.1.2) reduce to

$$x(t) = x_0\, e^{-(\theta+\sigma)t} + \xi\{1 - e^{-(\theta+\sigma)t}\},$$

$$y(t) = y_0\, e^{-(\theta+\sigma)t} + \eta\{1 - e^{-(\theta+\sigma)t}\}. \tag{6.2.1}$$

Consequently, one must have, for example, $\xi \geqslant 0$ in order that $x(t) > 0$ for all finite $t \geqslant 0$. Thus, we should anticipate having to make some sort of assumption about ξ and η in order to develop battle-outcome prediction conditions.

The first step in the development of these conditions is to generalize the nonoscillation of solutions to Lanchester-type equations for modern warfare to the case of the general linear model with constant attrition-rate coefficients and replacements (6.1) (cf. Taylor [18, Theorem 1] or Taylor [20, Result 4.1]).

THEOREM 2: Assume that $ab \neq \alpha\beta$. Then for all finite $t \geqslant 0$

$$\text{either} \quad x(t) > \xi \text{ or } y(t) > \eta. \tag{6.2.2}$$

If either of conditions (6.2.2) is violated at some finite time, then the other holds for all time $t \geqslant 0$.

PROOF: Recalling that the substitution (6.1.18) transforms (6.1.27) into (6.1.19) , we observe that Equations (6.1.19) are of the same mathematical form as those for the model without replacements and withdrawals (5.1). Consequently, we may apply Theorem 1 to (6.1.19) and conclude that $X(t)$ and $Y(t)$ between them have at most one zero for all finite time $t \geqslant 0$. In other words, either $X(t)$ or $Y(t) > 0$ for all finite $t \geqslant 0$, i.e., by (6.1.18) we have shown that (6.2.2) holds for all finite $t \geqslant 0$. Furthermore, if (for example) $X(t)$ is equal to zero at some finite time, we know from the proof of Theorem 1 that $Y(t) > 0$ for all $t \geqslant 0$, whence follows the last part of the theorem.

Q.E.D.

We can now develop rather precise battle-outcome prediction conditions if we assume that both ξ and $\eta \geqslant 0$.

THEOREM 3: Assume that $ab > \alpha\beta$ and that both ξ and $\eta \geqslant 0$. Then the X force will lose a fight to the finish in finite time (i.e., will be annihilated first in finite time) if and only if

$$(x_0 - \xi) < \left[\frac{\theta + \delta}{b}\right](y_0 - \eta), \tag{6.2.3}$$

which may also be written in the equivalent form

$$(x_0 - \xi) < \sqrt{R}\left\{\frac{S}{2} + \sqrt{\left[\frac{S}{2}\right]^2 + 1}\right\}(y_0 - \eta). \tag{6.2.4}$$

Furthermore, when (6.2.4) holds, we have that for all $t \geqslant 0$

$$y(t) > \eta \geqslant 0. \tag{6.2.5}$$

Neither side will be annihilated in finite time if an only if

$$(x_0 - \xi) = \sqrt{R}\left\{\frac{S}{2} + \sqrt{\left[\frac{S}{2}\right]^2 + 1}\right\}(y_0 - \eta), \tag{6.2.6}$$

and in this case

$$\lim_{t \to +\infty} x(t) = \xi \geqslant 0 \text{ and } \lim_{t \to +\infty} y(t) = \eta \geqslant 0. \tag{6.2.7}$$

PROOF: Sufficiency of (6.2.4) for the annihilation of the X force is proven by observing that $\theta - \sigma > 0$ if and only if $ab > \alpha\beta$. Consequently, we see from (6.1.1) that the X force level $x(t)$, which is a continuous function of time, becomes negative for t large enough when the coefficient A in (6.1.1) is negative, i.e., (6.2.3) holds. Hence, (6.2.3) holding [equivalently, (6.2.4)] guarantees that the X force will be annihilated in finite time, and then the last line of Theorem 2 guaranteees that (6.2.5) holds. Necessity of (6.2.3) for the annihilation of the X force (and also the statements concerning the nonannihilation of either side) is proven by observing from (6.1.1) and (6.1.2) that when $A = 0$, i.e., (6.2.6) holds, we have via (6.2.1) that $x(t)$ and $y(t) > 0$ for all finite $t \geqslant 0$. This result shows that X cannot be annihilated when $A = 0$ and proves the nonannihilation statements involving (6.2.6) and (6.2.7), since (6.2.1) is readily seen to yield (6.2.7). If $A > 0$, then (6.1.2) shows us that $y(t)$ has a finite zero point, since $\theta + \delta > 0$. Consequently, $y(t)$ can be made to be negative for t large enough, and Theorem 2 then yields that $x(t) > \xi \geqslant 0$ for all finite $t \geqslant 0$. Thus, we have shown that $x(t)$ can vanish only if $A < 0$, i.e., (6.2.4) is necessary for the X force to be annihilated in finite time.

Q.E.D.

Still assuming that both ξ and $\eta \geqslant 0$, we see that the Y force will be annihilated first in finite time if and only if the above inequality (6.2.4) is reversed. Furthermore, it should be clear that the assumption that both ξ and $\eta \geqslant 0$ plays an essential role in these battle-outcome prediction conditions. For example, the inequality (6.2.4) holding for the initial force levels

guarantees via Theorem 2 that $y(t) > 0$ for all $t \geqslant 0$ only when $\eta \geqslant 0$ as the numerical example depicted in Table I shows.* In this numerical example, the condition (6.2.4) is satisfied and we consequently still have that for all $t \geqslant 0$

$$y(t) > \eta, \tag{6.2.8}$$

but (6.2.8) no longer guarantees that the Y force will never be annihilated when $\eta < 0$. Thus, (6.2.4) being satisfied no longer guarantees that Y will win a fight to the finish (i.e., a battle that lasts until one side or the other has been annihilated) when $\eta < 0$. In other words, (6.2.4) is satisfied for the particular battle depicted in Table I, but the Y force is actually annihilated before the X force is. Again, the reason why (6.2.4) fails to correctly predict battle outcome is that $\eta < 0$. This numerical example should alert the reader to the fact that determination of the qualitative behavior, e.g., battle-outcome prediction, for the constant-coefficient model with a constant rate of replacement/withdrawal is in some sense much trickier for the variable-coefficient model with no replacements/withdrawals (5.1).

Moreover, we can relax somewhat our assumption concerning the signs of ξ and η and still develop sufficient condition for a side to win a fight to the finish.

THEOREM 4: Assume that $ab > \alpha\beta$ and that $\eta \geqslant 0$. Then the X force will lose a fight to the finish (i.e., a fight that lasts until the annihilation of one side or the other) in finite time if (6.2.4) holds. We also have for this battle that $y(t) > \eta \geqslant 0$ for all $t \geqslant 0$.

PROOF: Very similar to the proof of Theorem 3, since $A < 0$, we know from (6.1.1) that $x(t)$ becomes negative for t large enough, and hence $x(t)$ has a finite zero point. Since $x(t)$ can always be made to violate (6.2.2) by taking t large enough, Theorem 2 tells us that $y(t) > \eta \geqslant 0$ for all $t \geqslant 0$.

Q.E.D.

However, as the example depicted in Table I shows, "if-and-only-if" statements regarding battle outcome can no longer be made when we assume that only one of ξ and $\eta \geqslant 0$ (in the numerical example under consideration, only $\xi \geqslant 0$).

7. DISCUSSION

The results that we have given in this paper concerning the effects of replacements/withdrawals on battle outcome not only are of interest in their own right, but they are also important for understanding the behavior of complex operational Lanchester-type combat models that are widely used today in both the United States (e.g., see [2], [4], [19], [22]) and also NATO countries [8], [9]. In particular, a simple combat model such as we have considered here may yield a clearer understanding of important relations that are difficult to perceive in such a complex model. Furthermore, the mathematical combat model that we have studied here is important not only for the significance of particular modeling interpretations but also for its versatility (see Section 2 for further details). In particular, its interpretation as

*Here, we have computed $x(t)$ and $y(t)$ as functions of time from (6.1.1) and (6.1.2) and have ignored the nonnegativity restriction on them when the differential equations (6.1) are viewed as a Lanchester-type combat model. In actual application, however, one would terminate the battle when first $y = 0$. Moreover, in practice one decides to do this by computing the force-level trajectories $x(t)$ and $y(t)$ from (6.1.1) and (6.1.2). Thus, one is interested in these formal mathematical solutions without consideration of the physical restriction on negative force levels.

TABLE I—Example that Shows that One Must Have both ξ and $\eta \geqslant 0$ in Order for the Inequality (6.2.4) to Correctly Predict a Y Victory in a Fight-to-the-Finish.

t	$x(t)$	$y(t)$
0.00	200.00	60.00
0.1	172.26	33.31
0.2	151.52	13.73
0.3	135.94	−0.56
0.4	124.17	−10.92
0.5	115.19	−18.33
0.6	108.25	−23.53
0.7	102.79	−27.07
0.8	98.40	−29.35
0.9	94.76	−30.65
1.0	91.64	−31.18
1.1	88.85	−31.11
1.2	86.27	−30.53
1.3	83.78	−29.53
1.4	81.30	−28.15
1.5	78.76	−26.43
1.6	76.10	−24.37
1.7	73.27	−21.99
1.8	70.23	−19.28
1.9	66.92	−16.22
2.0	63.31	−12.79
2.1	59.36	−8.98
2.2	55.02	−4.73
2.3	50.23	−0.02
2.4	44.96	5.19
2.5	39.15	10.97
2.6	32.72	17.36
2.7	25.63	24.43
2.8	17.80	32.35
2.9	9.15	40.89
3.0	−0.41	50.44
3.1	−10.98	61.00
3.2	−22.66	72.67
3.3	−35.56	85.57
3.4	−49.82	99.82
3.5	−65.57	115.58

NOTE: In this battle we have taken (in compatible units) $a = b = 2$, $\alpha = \beta = 1$, $r = 0$, and $s = 150$. It follows that (6.2.4) is satisfied but with $\xi = 100$ and $\eta = -50$.

representing a simple model of large-scale combat over a sustained period of time with replacements continuously committed to battle generates particular interest within the framework of the NATO scenario with regard to rough insights into the structure of tradeoffs among the quality of weapon systems, the quantity of weapon systems, and the "build-up" rates at which new systems are introduced into a theater of operations.

Unfortunately, the general linear model with replacements/withdrawals (2.1) is much more difficult to analyze than its simpler versions (3.5) and (5.1) previously studied (e.g., see [16] for further details and references; also [17], [20]). We show how the general linear model (2.1) could be transformed to a simpler canonical form (3.2), but unfortunately this canonical form has not been much studied. We did see, however, that a fixed-force-level-breakpoint battle modeled with Lanchester-type equations for modern warfare (3.5) does lead to this same canonical form. More importantly, solutions to the canonical form (3.2) were shown to exhibit qualitative behavior (at least in special cases) markedly different from that for Lanchester-type equations of modern warfare (3.5). In particular, nonoscillation in the strict sense was shown not to hold for solutions to (at least the constant-coefficient version of) (2.1).

We showed how to represent the force levels for the model (2.1) in terms of general Lanchester functions corresponding to the model without replacements/withdrawals, but these analytical results do not appear to be of much practical significance for developing insights into the dynamics of combat. More importantly, we showed in Theorem 1 that all solutions to this model without replacements/withdrawals had the property of being nonoscillatory (in the strict sense) in marked contrast to that of those to the general model (2.1) (recall the example depicted in Table I). This analysis of solutions to (5.1) turned to be be particularly significant when we saw that it could be applied to the constant-coefficient version of (2.1).

In hopes of developing insights into the dynamics of combat, we turned to the general linear model with constant attrition-rate coefficients and replacements (6.1) because of the apparent opacity of analytical results for its variable-coefficient version. Unfortunately, analytical results for the force levels to (6.1) turned out to be so complex (although wholly and explicitly expressible in terms of elementary functions, i.e., exponentials) that we were unable to determine the qualitative behavior of the force-level trajectories by direct examination of these analytical results. However, we were able in Theorem 2 to generalize the nonoscillation of solutions to this model by transforming its equations (6.1) into those for combat with only supporting fires (5.1) [recall Equations (6.1.19)]. Theorem 2 was particularly significant because we were able to develop the battle-outcome prediction conditions of Theorems 3 and 4 by using it.

The complexity of solution behavior for the constant-coefficient model (6.1) was apparently not appreciated in the earlier work of Morse and Kimball [14], and consequently incorrect battle-outcome prediction conditions were implied in [14, p. 72] (recall the example depicted in Table I). Because of our previous theoretical examinations (particularly with respect to the nonoscillation of solutions) of the simpler models (3.5) and (5.1), we had developed enough analytical machinery to profitably study the more complex model at hand (2.1) [especially its constant-coefficient version (6.1)]. Such results were, of course, not available when Morse and Kimball wrote their pioneering book on military operations research [14]. Finally, the reader should note that we have shown how various Lanchester-type combat models are transformable into each other and consequently how results for simpler models can be adapted to the study of more complicated ones.

8. CONCLUSIONS

(I) Analytical results for Lanchester-type models with continuous replacements and/or withdrawals are much more complicated than those for models without replacements/withdrawals. The complexity of these analytical results is such that general results concerning the qualitative behavior of solutions in general is required for the study of theier qualitative behavior in specific instances.

(II) The problem of studying a fixed-force-level-breakpoint battle modeled with Lanchester-type equations for modern warfare is mathematically equivalent to that of studying a fight to the finish with continuous withdrawals on both sides.

(III) The force-level trajectories for the so-called general linear model with continuous replacements/withdrawals behave in a fundamentally different way (at least in special cases such as constant coefficients) than do those for the corresponding model without replacements/withdrawals (i.e., so-called combat with (only) supporting fires). In particular, for the general linear model with continuous replacements, both sides may successively be annihilated in battle if the force-on-force attrition equations are not "turned off" at the right time, and this phenomenon makes battle-outcome prediction particularly difficult.

REFERENCES

[1] Bach, R.E., L. Dolansky, and H.L. Stubbs, "Some Recent Contributions to the Lanchester Theory of Combat," Operations Research, *10*, 314-326 (1962).

[2] Berg, D.J. and M.E. Strickland, "Catalog of War Gaming and Military Simulation Models (7th Edition)," SAGA-180-77, Studies, Analysis, and Gaming Agency Organization of the Joint Chiefs of Staff, Washington, D.C., (August 1977).

[3] Brewer, G.D., "Analysis of Complex Systems: An Experiment and Its Implications for Policymaking," P-4951, The RAND Corporation, Santa Monica, CA, (January 1973).

[4] Command and Control Technical Center, "VECTOR-2 System for Simulation of Theater-Level Combat," TM 201-79, Washington, D.C., (January 1979).

[5] Dare, D.P., "On a Hierarchy of Models," in *Operationsanalytische Spiele für die Verteidigung/Operational Research Games for Defense*, 285-307, R.K. Huber, K. Niemeyer, and H.W. Hofmann, Editors, (R. Oldenbourg Verlag, München, Germany, 1979).

[6] Geoffrion, A.M., "The Purpose of Mathematical Programming is Insight, Not Numbers," INTERFACES *7*, 1, 81-92 (1976).

[7] Hildebrand, F.B., *Advanced Calculus for Engineers*, (Prentice-Hall, Englewood Cliffs, NJ, 1948).

[8] Huber, R.K., L.F. Jones, and E. Reine, Editors, *Military Strategy and Tactics*, (Plenum, NY, 1975).

[9] Huber, R.K., K. Niemeyer, and H.W. Hofmann, Editors, *Operationsanalytische Spiele für* die Verteidigung/Operational Research Games for Defense, (R. Oldenbourg Verlag, München, Germany, 1979).

[10] Ignall, E.J., P. Kolesar, and W.E. Walker, "Using Simulation to Develop and Validate Analytic Models: Some Case Studies, "Operations Research *26*, 237-253 (1978).

[11] Karns, C.W., "An Application of Lanchester's Equations to Amphibious Assaults," ORG Study No. 1, ORG 126-53, Operations Research Group, Office of Naval Research, (December 1953) (AD 38 491).

[12] Lanchester, F.W., "Aircraft in Warfare: The Dawn of the Fourth Arm—No. V., The Principle of Concentration," Engineering 98, 422-423 (1914). (Reprinted on pp. 2138-2148 of The World of Mathematics, Vol. IV, J. Newman, Editor, (Simon and Schuster, NY, 1956).

[13] Morris, W.T., "On the Art of Modelling," Management Science, *13*, B-707-B-717 (1967).

[14] Morse, P.M. and G.E. Kimball, *Methods of Operations Research*, (The M.I.T. Press, Cambridge, MA, 1951).

[15] Stockfisch, J.A., "Models, Data, and War: A Critique of the Study of Conventional Forces," R-1526-PR, the RAND Corporation, Santa Monica,CA, (March 1975).

[16] Taylor, J.G., "Recent Developments in the Lanchester Theory of Combat," in *Operational Research '78, Proceedings of the Eighth IFORS International Conference on Operational Research*, 773-806, K.B. Haley, Editor, (North-Holland, Amsterdam, 1979).

[17] Taylor, J.G., "Some Simple Victory-Prediction Conditions for Lanchester-Type Combat Between Two Homogeneous Forces with Supporting Fires," Naval Research Logistics Quarterly. *26*, 365-375 (1979).

[18] Taylor, J.G., "Prediction of Zero Points of Solutions to Lanchester-Type Differential Combat Equations for Modern Warfare," Journal of Applied Mathematics *36*, 438-456 (1979).

[19] Taylor, J.G., *Force-on-Force Attrition Modelling*, Military Applications Section of the Operations Research Society of America, Arlington, VA, (1980).

[20] Taylor, J.G., "Theoretical Analysis of Lanchester-Type Combat Between Two Homogeneous Forces with Supporting Fires," Naval Research Logistics Quarterly, 109-121 (1980).

[21] Taylor, J.G. and S.H. Parry, "Force-Ratio Considerations for Some Lanchester-Type Models of Warfare," Operations Research, *23*, 522-533 (1975).

[22] U.S. General Accounting Office, "Models, Data, and War: A Critique of the Foundation for Defense Analyses," PAD-80-21, Washington, D.C., (March 1980).

THE POWER OF PAIRED SAMPLE TESTS

Stuart Jay Deutsch
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia

Bruce Wayne Schmeiser

Purdue University
West Lafayette, Indiana

ABSTRACT

The power of the component randomization test for the paired sample location test is compared to the power of the parametric test for paired comparisons using the sum of differences distribution. This analysis is presented in the form of operating characteristic curves for normal, exponential, uniform and absolute lambda observations over a range of small sample sizes.

1. INTRODUCTION

Little study of the power of component randomization tests for paired comparisons has been undertaken for finite sample sizes. Knowledge of the power is important for evaluating the tradeoffs between the robustness and computational ease of the various tests. The component randomization test, the most powerful distribution-free test, has the disadvantage of requiring great computational effort. The corresponding parametric test, which assumes knowledge of the exact form of the distribution of the test statistic, would be expected to have more power than distribution-free tests which do not make use of this information, but knowledge of this distribution is rare. The normal test is of interest because it is widely used and uniformly most powerful in the case of normally distributed observations.

Quantifying the difference in power between the first two tests is helpful in deciding which test to use in a particular situation. The power of the normal test under the null hypothesis, which is the probability of type I error, is important in determining the degree of inappropriateness of the normal test for various underlying distributions.

In Deutsch and Schmeiser [1], [2], methods were developed to study the component randomization test and the exact corresponding parametric test, respectively. The methods are general in that they are appropriate for any underlying distribution and sample size. These techniques are implemented here, using also the methods of quantile estimation discussed in Deutsch and Schmeiser [3] and of process generation discussed in Schmeiser and Deutsch [4], to estimate the power of various tests of location for the paired sample design.

Specifically, the tests considered are the component randomization test, the corresponding parametric test, and the commonly used normal test. The "corresponding parametric test" is to reject H_0 if the test statistic is greater than the $(1-\alpha)$ quantile of the distribution of the test statistic. In the case of normally distributed observations, the parametric test and the normal test are one and the same.

In the next section, an overview of the analysis performed for each underlying distribution and sample size in later sections is described. In Section 3 an analysis of the power of these tests under several specific distributions and sample sizes is presented. Lastly, in Section 4, conclusions of the analyses are delineated.

2. ANALYSIS METHODOLOGY

Before looking at specific results, the steps necessary to determine the powers of the tests of interest are now discussed. Conceptually, the approach is straightforward, requiring only two steps:

(a) Determine $F_n^{-1}(p)$, the pth quantile of the reference distribution F for the parametric test. The power of the parametric test or the normal test may then be determined from $1 - \beta = 1 - F(C_\alpha - \sqrt{n}E\{d_i/\sigma\})$ where C_α is the $(1-\alpha)$ quantile of F or the normal distribution.

(b) Generate N samples of size n from the distribution of interest with specified expected differences. For each sample perform the four tests and update a counter for each test which rejects H_0. The count divided by N is then the estimate of the power of the test.

Thus the power of the parametric tests may be determined from either phase. Note that the power of the component randomization test may be determined only from Phase 2.

3. THE POWER OF CERTAIN TESTS

In this section the results arising from the techniques developed previously are presented. The probability of type II error, β, which is one minus the power, is given in both tabular and graphical form for each of the tests for various underlying distributions and sample sizes. The aim of this section is to examine some specific examples to determine the relationship of power to other factors in a general way. The normal, exponential, uniform, and absolute lambda cases are examined. These distributions were selected to illustrate the effect of tail weight and shape on the power of the tests.

In each section results are given in the form of operating characteristic (O.C.) curves. The curves for sample sizes 4, 7, 11, and 15 are grouped by underlying distribution and α value. the values of α considered are 1/128, 1/16, and 1/8, the closest values to .01, .05, and .1 of the form $m/2^n$ where n is integer. These values are necessary since α can take on only values of this form for permutation tests.

Normal Observations

In this section the power of the tests of interest, presented in the form of O.C. curves, are compared for the case of normally distributed observations. Since the appropriate parametric test is the normal test, only the component randomization and normal tests are of interest here.

Figure 1 shows graphically the results of this section, which are also presented in tabular form in Table 1. Several observations can be made from the graphs corresponding to $\alpha = 1/128$, $1/16$, and $1/8$, respectively:

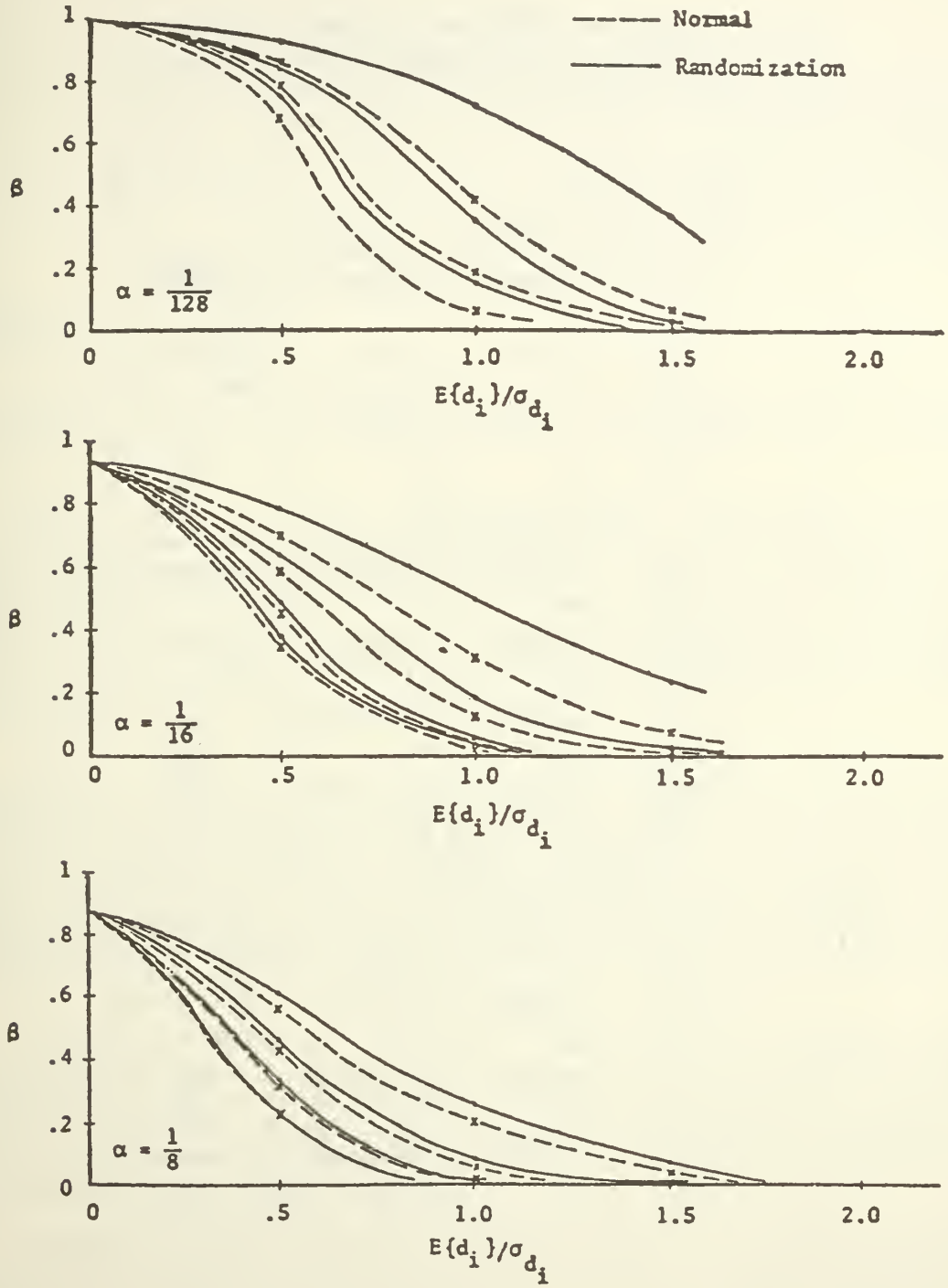


FIGURE 1. Operating characteristic curves from normal observations.

TABLE 1 — *Type II Errors (β) of Randomization and Parametric Test (Standard Deviations Shown in Parentheses)*

δ/σ n	0	.5	1.0	1.5
	$X, Y \sim N(0,.5)$ $\alpha = 1/128 = .0078125$ $z_\alpha = 2.417$			
7	.991875 .991875	.93(.0046) .86	.71(.0083) .41	.38(.0088) .06
11	.991875 .991875	.84(.0081) .78	.35(.015) .18	.036(.0059) .005
15	.991875 .991875	.76 (.017) .69	.15(.013) .07	neg neg
	$\alpha = 1/16 = .0625$ $z_\alpha = 1.534$			
4	.9375 .9375	.77(.0066) .70	.50(.0079) .32	.24(.0067) .07
7	.9375 .9375	.64(.0088) .58	.19(.0072) .13	.02(.0026) .01
11	.9375 .9375	.49(.011) .45	.051(.0070) .04	neg neg
15	.9375 .9375	.38(.020) .35	.012(.0040) .01	neg neg
	$\alpha = 1/8 = .125$ $z_\alpha = 1.150$			
4	.875 .875	.61(.0077) .56	.27(.0069) .20	.075(.0042) .03
7	.875 .875	.46(.0091) .43	.09(.0051) .07	.005(.0013) .002
11	.875 .875	.33(.010) .31	.016(.0040) .015	neg neg
15	.875 .875	.22(.017) .22	.0038(.0022) .003	neg neg

1. The value of β decreases as sample size increases.

2. The value of β decreases as $E\{d_i/\sigma\}$ increases.

3. The value of β decreases as α increases.

4. The normal test, which dominates the component randomization test in this case of normally distributed observations, has asymptotically the same power as the component randomization test. The dominance of the normal test decreases as α increases and as n increases. This is logical since many observations are needed for the tail areas of the randomization reference distribution to be well developed. An α value far out in the tail requires more observations for the same power. Thus, for normally distributed observations, the largest loss of power in using the component randomization test is for small n and small α.

Exponential Observations

Results are given here for exponentially distributed observations. The form of the results is the same as for the normal case of the last section, except here the parametric test is distinct from the normal test.

Figure 2 shows the parametric reference distribution for the parametric test under exponential observations as determined by Phase 1. The upper half of the reference distribution is plotted on normal probability paper for sample sizes $n = 4, 7, 11, 15$, and infinity. The reference distribution as n approaches infinity is the normal distribution (the straight line in the figure) with the difference between the normal distribution and the parametric reference distribution for finite sample sizes being greatest for small samples. For all sample sizes the normal distribution lies above the parametric reference distributions above approximately the 95% quantile and below the parametric reference distribution otherwise.

Table 2 contains the results of this section in tabular form. Three tables are shown, corresponding to $\alpha = 1/128$, $1/16$, and $1/8$, respectively. Each of the entries are the probability of type II error, β, for various values of n and $E\{d_i/\sigma\}$. Results for the component randomization test are in the columns labeled "R." Similarly the parametric test results are under "P" and the normal test results are under "N." For the randomization test the estimate of the standard deviation of the result is given in parentheses. The results for "P" and "N" are deterministic calculations from the results of Phase 1. The critical values used for the parametric test are determined from Figure 2 and are given in the right-hand column under C_α.

The most surprising aspect of these results is that the power of the parametric test does not dominate the power of the component randomization test. Especially for alternative hypotheses close to H_0, the component randomization test has more power for all n and values of α. At first glance this is counter-intuitive, since the usual circumstance is that power is lost in obtaining distribution-free properties. Note that over the range of values of $E\{d_i/\sigma\}$, however, the parametric test is indeed more powerful.

An intuitive rationale for the greater power of the component randomization test for alternative hypotheses close to the null hypothesis is as follows: the reference distribution for

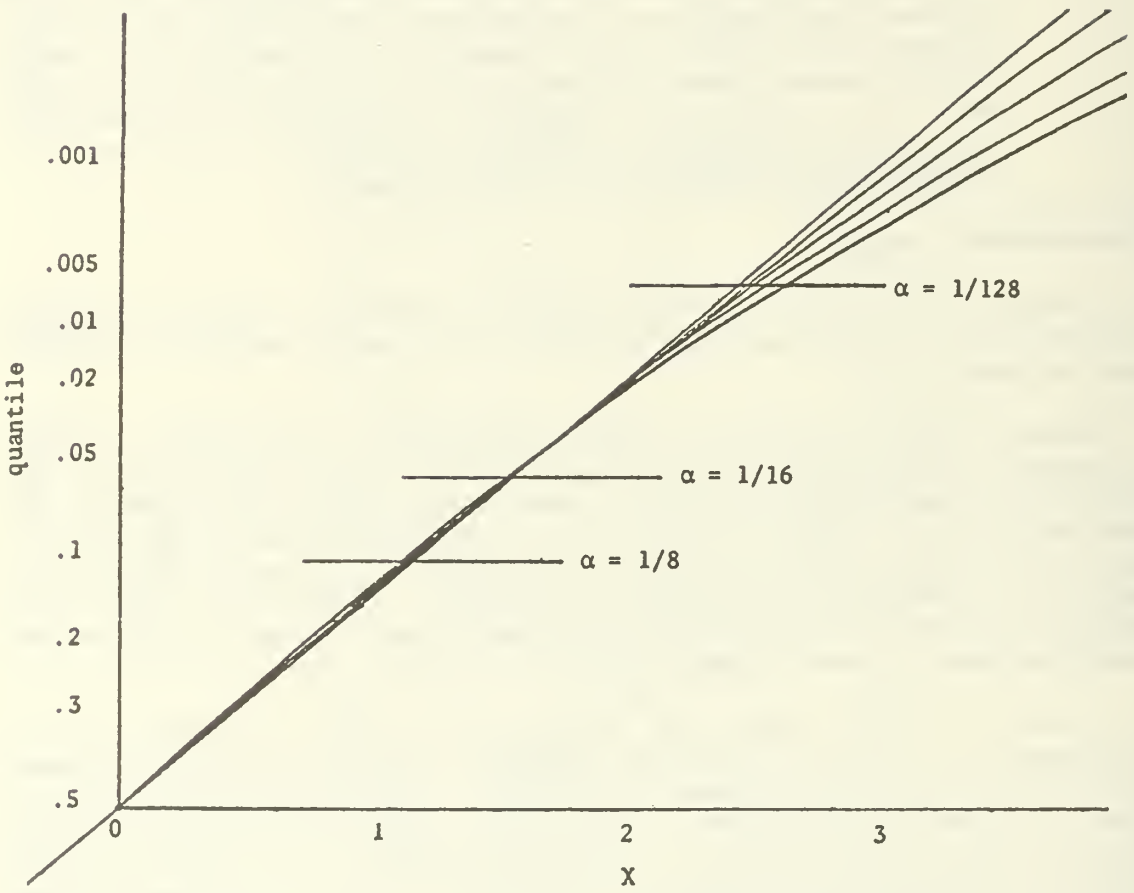


FIGURE 2. Reference distributions arising from exponential observations.

TABLE 2 — *Type II Errors (β) for Randomization (R), Parametric (P), and Normal (N) Tests Arising from Exponential Observations*

δ D	0			.5			1.0			1.5			C_α
n	R	P	N	R	P	N	R	P	N	R	P	N	
	$X, Y \sim$ Exp $\alpha = 1/128 = .0078125$ $z_\alpha = 2.417$												
7	$\frac{127}{128}$	$\frac{127}{128}$	.990	.86 (.0063)	.89	.86	.59 (.0090)	.48	.40	.34 (.0086)	.07	.06	2.54 (.02)
11	$\frac{127}{128}$	$\frac{127}{128}$	.990	.76 (.013)	.80	.79	.30 (.0089)	.20	.18	.07 (.0046)	.008	.007	2.49 (.01)
15	$\frac{127}{128}$	$\frac{127}{128}$	.991	.66 (.024)	.70	.69	.13 (.023)	.08	.07	neg	neg	neg	2.47 (.03)
	$\alpha = 1/16 = .0625 z_\alpha = 1.534$												
4	.9375	.9375	.940	.67 (.0074)	.70	.69	.40 (.0077)	.30	.31	.21 (.0065)	.06	.07	1.51 (.004)
7	.9375	.9375	.939	.54 (.0091)	.61	.61	.17 (.0069)	.12	.13	.035 (.0034)	.01	.01	1.52 (.007)
11	.9375	.9375	.938	.42 (.016)	.45	.45	.0.68 (.0046)	.04	.04	.005 (.0013)	.0005	.0005	1.53 (.004)
15	.9375	.9375	.938	.34 (.019)	.34	.34	.01 (.007)	.01	.01	neg	neg	neg	1.53 (.02)
	$\alpha = 1/8 = 1.25 z_\alpha = 1.150$												
4	.875	.875	.89	.51 (.0079)	.54	.56	.22 (.0065)	.16	.18	.08 (.0043)	.03	.035	1.09 (.004)
7	.875	.875	.88	.40 (.0090)	.43	.41	.09 (.0052)	.06	.07	.013 (.002)	.005	.005	1.11 (.006)
11	.875	.875	.88	.29 (.014)	.30	.30	.03 (.0031)	.01	.01	.002 (.0008)	neg	neg	1.13 (.006)
15	.875	.875	.88	.20 (.016)	.20	.20	neg	neg	neg	neg	neg	neg	1.14 (.01)

exponential observations has heavier tails than the normal ($\alpha_4 = 3.75$ for $n = 4$, and the normal value is $\alpha_4 = 3$) while the randomization reference distribution has light tails ($\alpha_4 < 3$) since its range is always finite. A distribution with a high value of α_4 is more peaked and has heavier tails than a distribution with smaller α_4. Small α_4 values indicate light tails and heavy shoulders. Now an alternative hypothesis close to the null hypothesis is more easily detected by the distribution with heavy shoulders, in this case the randomization reference distribution, since many observations are close to the mean. On the other hand, alternative hypotheses in the tails are more easily detected by the parametric test due to its reference distribution having more observations in the tails.

Note that the power of the normal theory test is very similar to the parametric test. While not appropriate for exponential observations, the nominal value of α is not far from the true value, as indicated in the columns under $E\{d_i/\sigma\} = 0$. The power of the normal test is similar to the power of the parametric test for all alternative hypotheses and sample sizes, although the greatest difference is for small sample sizes where the parametric reference distribution differs the most from the normal distribution.

Note also that as the sample size increases the results of all three tests converge to the same values. This must be true for the parametric and normal tests due to the central limit theorem. This is equivalent to the reference distributions becoming the same in Figure 2 for large sample sizes. The component randomization test also converges to the same power for large sample sizes, since this test is asymptotically the normal theory test for large sample sizes. However, the nonnormality in terms of sample size is of interest here.

Figure 3 shows graphically the values of Table 2 for the parametric and component randomization tests. The normal test is not shown since it is only approximately correct for this case and since the results are so similar to the parametric test. All of the same general relationships between α, n, and $E\{d_i/\sigma\}$ are true here as for the normal case. β decreases as any of the three factors increase. The striking difference is that the parametric test does not dominate the component randomization test, as just discussed.

It is of interest to compare the results of this section to those of the last section for normal observations. The power of the component randomization test is generally greater for exponential observations than for normal observations, the greatest difference being .1 for $n = 4$, $\alpha = 1/16$, and $E\{d_i/\sigma\} = .5$. However, for large values of $E\{d_i/\sigma\}$ and/or for large n the normal observations lead to the greater power. In these cases the power is so great for both tests, however, that the difference is not usually important. Thus it appears that the component randomization test actually has better overall power for exponential observations than for normal observations.

This relationship does not hold true for the parametric test. For small α, the parametric test performs best under normal observations. For $\alpha = 1/16$ the test has similar power for both types of observations and for large α values the parametric test performs best for exponential observations. That this should be true is obvious from Figure 2. The reference distributions all cross the normal distribution around the .95 quantile. Since power of the parametric test depends on this distribution only through the $(1\text{-}\alpha)$ quantile, power should indeed be similar for $\alpha = 1/16$.

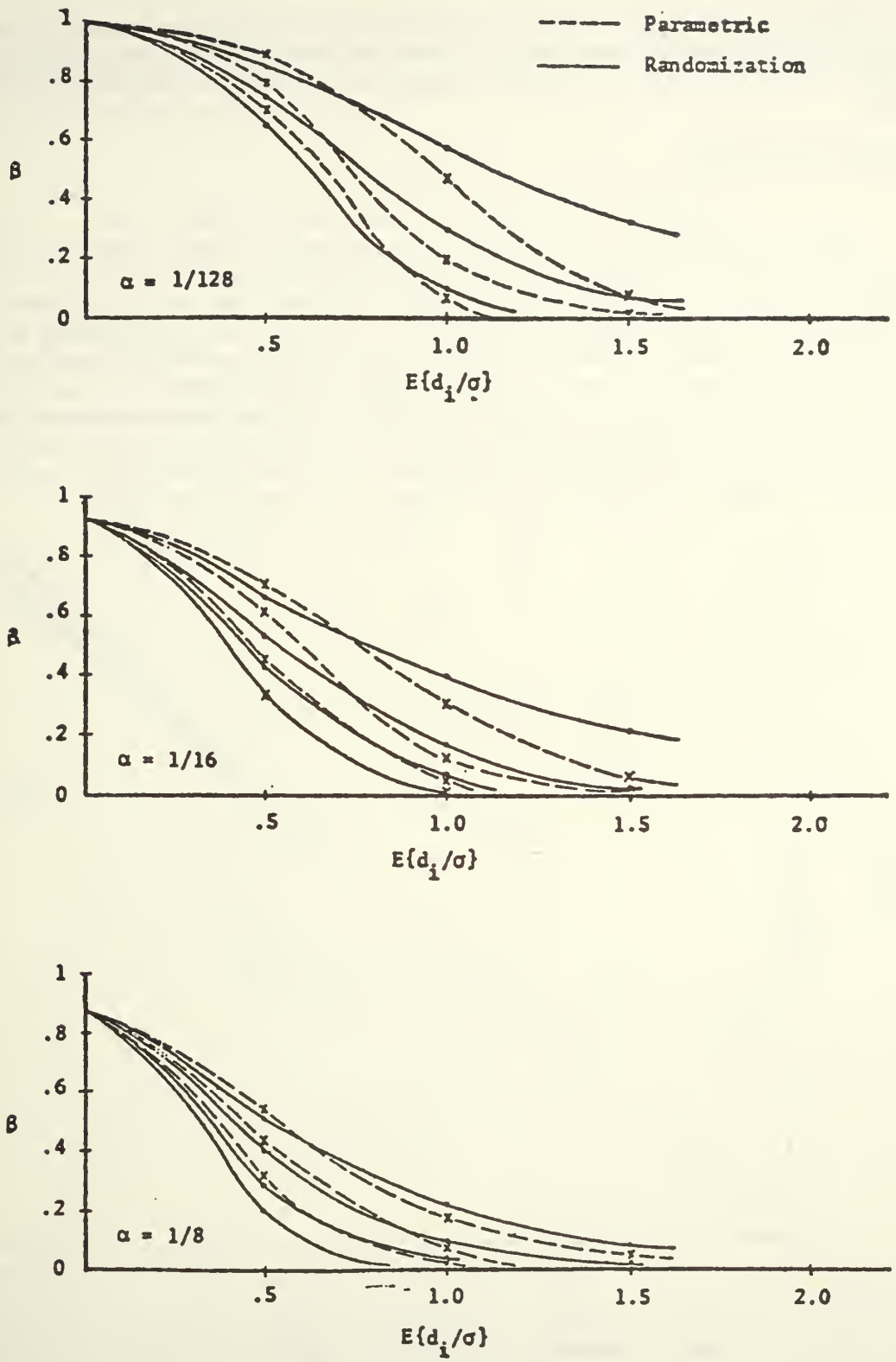


FIGURE 3. Operating characteristic curves from exponential observations.

Note that the result of all the parametric reference distributions having the same .95 quantile is that the normal theory test is very close to exactly valid for exponential observations for any sample size if $\alpha \simeq .05$. Smaller or larger values of α lead to inaccuracies in the normal test.

Uniform Observations

Results corresponding to those of the last section are given here for uniformly distributed observations. Figure 4 shows the reference distributions for sample sizes $n = 4, 7, 11, 15$, and infinity arising from uniform observations. Again the straight line, the normal distribution, is the limiting distribution as n approaches infinity. Not shown is the limiting case of $n = 1$, the distribution of the difference between two uniformly distributed random variables, which is the triangular distribution. All of these reference distributions have a finite upper bound and therefore lie above the normal line, whereas the exponential reference distributions were below the normal distribution. Note that, for the α values considered, the reference distribution is essentially identical with the normal. Only for α greater than .01 is the difference noticeable in Figure 4 and even there it is not as great as in the case of exponential observations.

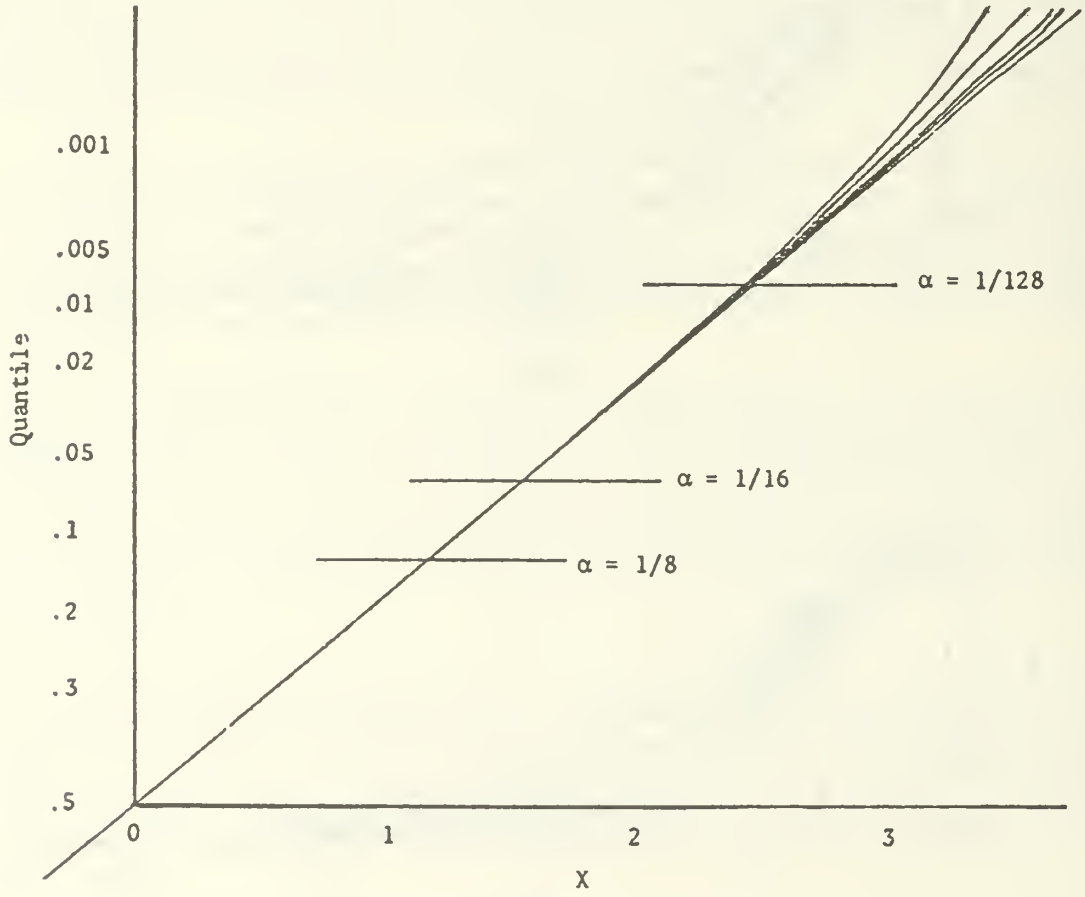


FIGURE 4. Reference distributions arising from uniform observations.

Table 3 shows the probability of type II error for various sample sizes and alternative hypotheses $E\{d_i/\sigma\}$ for the component randomization and parametric tests. From the table it would appear that the normal test works well in this case as an approximation to the true parametric test, even for small sample sizes.

TABLE 3 — *Type II Errors* (β) for Randomization (R), Parametric (P), and Normal (N) Tests Arising from Uniform Observations

δ D / n	0			.5			1.0			1.5			C_α
	R	P	N	R	P	N	R	P	N	R	P	N	
	$X, \tilde{Y}$ Uniform												
	$\alpha = 1/128 = .0078125$ $z_\alpha = 2.417$												
7	$\frac{127}{128}$	$\frac{127}{128}$	.99	.93 (.004)	.85	.85	.74 (.008)	.40	.41	.42 (.009)	.06	.06	2.39
11	$\frac{127}{128}$	$\frac{127}{128}$	.99	.87 (.01)	.80 (.01)	.80 (.015)	.40	.19 (.12)	.19	.03 (.005)	.004 (.002)	.004	2.40
15	$\frac{127}{128}$	$\frac{127}{128}$	.99	.81 (.03)	.67 (.04)	.67		neg			neg		2.41
	$\alpha = 1/16 = .0625$ $z_\alpha = 1.534$												
4	.9375	.9375	.94	.68 (.006)	.70	.70	.54 (.007)	.32 (.007)	.32	.27 (.006)	.07 (.003)	.07	1.53
7	.9375	.9375	.96	.65 (.009)	.57	.58	.21 (.007)	.14	.14	.02 (.002)	.005	.006	1.53
11	.9375	.9375	.96	.50 (.02)	.45 (.02)	.45	.05 (.006)	.04 (.005)			neg		1.53
15	.9375	.9375	.94	.34 (.04)	.29 (.04)	.29		neg			neg		1.53
	$\alpha = 1/8 = .125$ $z_\alpha = 1.150$												
4	.875	.875	.87	.52 (.006)	.56 (.006)	.56	.30 (.006)	.19 (.004)	.19	.08 (.003)	.03	.03	1.15
7	.875	.875	.87	.40 (.009)	.43	.43	.09 (.005)	.07	.07	neg	neg	neg	1.15
11	.875	.875	.87	.32 (.01)	.30 (.01)	.30	.01 (.004)	.01 (.004)			neg		1.15 1.15
15	.875	.875	.87	.19 (.03)	.18 (.03)	.18		neg			neg		

The results for uniform observations are similar to those of the prior two cases, with β decreasing as a function of n, α, and $E\{d_i/\sigma\}$. Again of interest is that the component randomization test has greater power (lesser β) for $E\{d_i/\sigma\} = .5$ and $\alpha = 1/16$ and $\alpha = 1/8$ than the parametric test. That is, even though the parametric test has better power for most combinations of n, α, and $E\{d_i/\sigma\}$, it does not dominate the component randomization test.

It is apparent by comparing Tables 1 and 2 that the normal test does indeed approximate the properties of the parametric test well in this case. The power never differs by more than .01 except for large values of $E\{d_i/\sigma\}$ where the difference in tail weights between the uniform and exponential distributions has an effect. If, however, a value of .0001 and a small sample size were used, Figure 4 shows that the normal test is not so good an approximation since $c_{.9999}$, equal to 3.35 for $n = 4$, does differ from the normal value $z_{.9999} = 3.45$, resulting in a biased indication of the true α value.

Absolute Lambda Observations

In the previous three subsections, the power of the component randomization test and the corresponding parametric test has been examined for normal, exponential, and uniform observations. These three distributions were selected as three distributions representing a wide range of tail weights. The uniform has the lightest tails ($\beta_2 = 1.8$), the normal has medium tails ($\beta_2 = 3$), and the exponential distribution has heavy tails ($\beta_2 = 9$). A question of interest is whether tail weight, as measured by the fourth standardized moment, is really the central factor in determining the effect of underlying distribution on the power of the tests.

An analysis similar to the last three sections is performed here to gain insight into the fourth moment's impact on power. In particular, observations are generated from the absolute lambda distribution (Deutsch and Schmeiser [4], with moments one through four identical to the standardized normal distribution. Despite having the same skewness and kurtosis, this density function has a shape quite unlike the normal.

Figure 5 shows the reference distribution arising from the aboslute lambda distribution with parameters $\lambda_1 = 0$, $\lambda_2 = -1.575$, $\lambda_3 = -.2247$, and $\lambda_4 = .5$ for a sample size of four. The reference distributions for larger sample sizes are not shown since in the figure they would not be distinguishable. Also included, for comparitive purposes, is the normal distribution which appears as a straight line. While the difference between the reference distribution shown and the normal is slight, compared to previous examples, it is significant. The major difference between this and previous examples is that the nonnormality occurs in the body, rather than the tail, of the distribution.

Table 4 shows the results of the Monte Carlo determination of the power of the component randomization and the parametric tests for this underlying distribution. The results for the normal test and the parametric test are identical to two places of accuracy, and therefore the results are given only for the parametric test. For the cases studied, the power of the test does not differ noticeably from the results for normally distributed observations. Note, however, that the power of the component randomization test is somewhat less for these observations than for normal observations. The discontinuity of the underlying distribution seems to affect the component randomization test while not affecting the parametric test.

4. CONCLUSIONS

The power of the component randomization test and the parametric test is given in the form of operating characteristic curves for normal, exponential, uniform, and absolute lambda observations. The power of the component randomization test is usually less than the power of the appropriate parametric test, but is *not* dominated for all alternative hypotheses. That is, in some cases, the distribution-free randomization test has more power than the parametric test which assumes knowledge of the exact distribution of the observations, including variance.

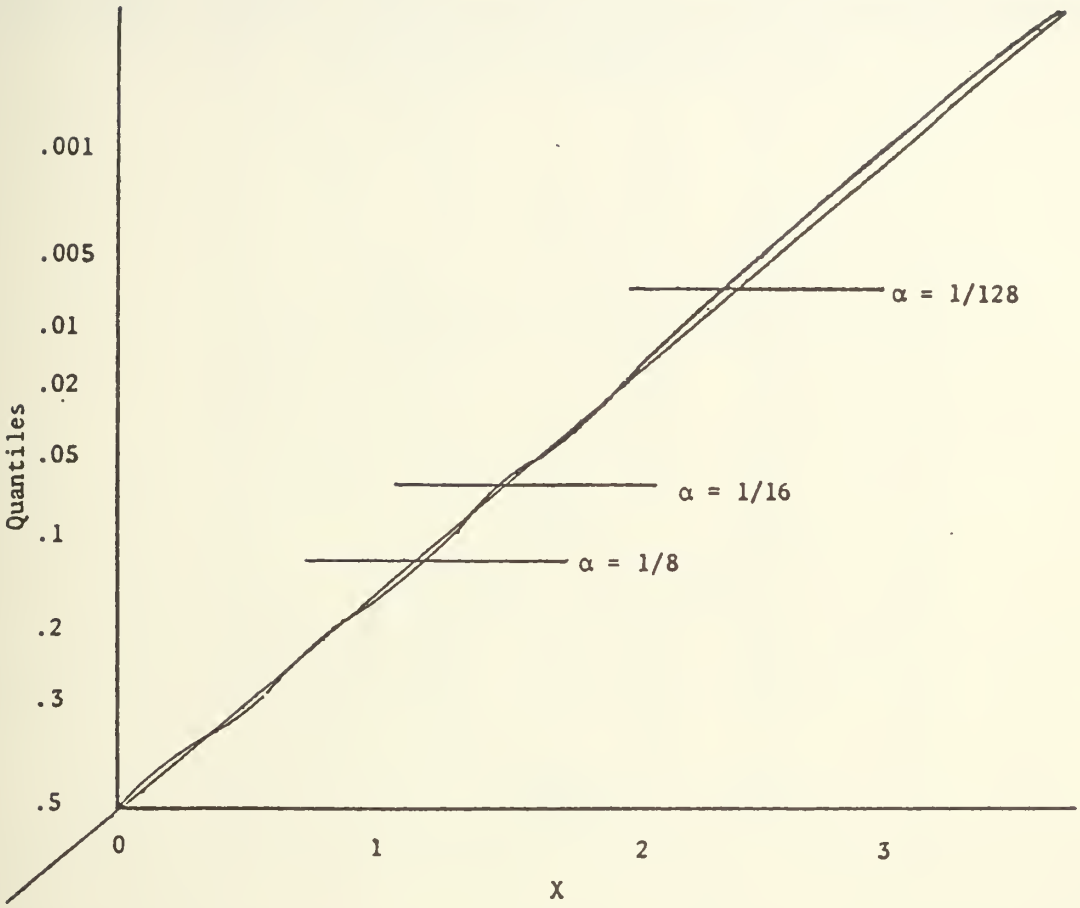


FIGURE 5. The reference distribution arising from absolute lambda observations for $n = 4$.

TABLE 4 — *Type II Errors (β) for Randomization (R), Parametric (P), and Normal (N)* Tests Arising from Absolute Lambda Observations (0, 2.227, −.224745, .5)

δ/σ	.5		1.0		1.5	
n	*R*	*P*	*R*	*P*	*R*	*P*
$\alpha = 1/128 = .0078125$			$z_\alpha = 2.417$			
7	.89	.87	.87	.42	.30	.06
11	.87	.78	.36	.19	neg	
15	.72 (.036)	.68	.14 (.028)	.08	neg	
$\alpha = 1/16 = .0625$			$z_\alpha = 1.534$			
4	.71	.67	.68	.32	.18	.06
7	.66	.59	.21	.14	.013	.007
11	.52	.47	.04	.03	neg	
15	.37 (.039)	.34	.013 (.009)	.013	neg	
$\alpha = 1/8 = .125$			$z_\alpha = 1.150$			
4	.66	.58	.29	.20	.06	.03
7	.49	.45	.09	.07	.003	.001
11	.33	.32	.01	.01	neg	
15	.29 (.035)	.23	.006 (.006)	0	neg	

The robustness of the normal test of location, a well-known property, was observed in relation to the performance of the correct parametric test. The difference in reference distributions is seen to be greatest in the tails of the distributions for common distributions. The power of the component randomization test is different for normal and absolute observations, even though both have identical first four moments. The power of the parametric test, on the other hand, is essentially the same for both types of observations.

Since in practice the appropriate parametric test is not known, the component randomization test is a viable alternative to the normal test for samples no larger than 25. Especially when the variance is unknown the component randomization test costs little in terms of power loss.

REFERENCES

[1] Deutsch, S.J. and B.W. Schmeiser, "The Computation of the Component Randomization Test for Paired Comparisons," Journal of Quality Technology, *15*, 2, 1983.
[2] Deutsch, S.J. and B.W. Schmeiser, "Estimation of the Sum of Differences Distribution," Communications in Statistics, *9*, 6, 1980.
[3] Schmeiser, B.W. and S.J. Deutsch, "Quantile Estimation from Grouped Data," Communications in Statistics, *6*, 3, 1977.
[4] Schmeiser, B.W. and S.J. Deutsch, "Versatile Four Parameter Family of Probability Distributions Suitable for Simulation," AIIE Transactions, *9*, 2, 1977.

EQUIVALENCE RELATIONS IN THE THEORY OF REPETITIVE CONSUMPTION

Joseph L. Balintfy

University of Massachusetts
Amherst, Massachusetts

Emmanuel S. Melachrinoudis

Northeastern University
Boston, Massachusetts

ABSTRACT

Preference for food in the function of the time since last consumption and total preference in the function of eating frequency are equivalent mathematical representations of attitudes governing food consumption. The observed and postulated properties of these functions imply a formal correspondence between preference maximizing and variety seeking behavior.

INTRODUCTION

Repetitive consumption is defined here as the phenomenon of consuming identical goods repeatedly over time. Repeated demand for identical goods may stem from technological necessity such as replacement, or from the desire to obtain the sensory stimulus associated with the consumption of known commodities. This article is concerned with the latter case.

Food consumption is the most prevalent example of repetitive consumption, although the same phenomenon may be manifested in "consuming" music, sights, or even certain fragrances. The common characteristic of these phenomena is that the act of consumption tends to create a feeling of satiation that inhibits the desire for further stimulus of the same kind for at least a short time period. As time progresses, however, the desire may build up again, and the consumption-satiation sequence is repeated.

The mathematical representation of this process is possible through the formulation of a preference function over time which expresses the desire for consuming a particular food item during the consumption-satiation cycle. The evidence reported by Balintfy, Duffy and Sinha [1] indeed indicates that such a preference-time function exists. Responses to questionnaires indicated that preference, i.e., the indicator of anticipated satisfaction, increases monotonically as a concave function of time since last consumption.

Another aspect of this phenomenon is the frequency of consumption for food items and people's ability to estimate their desired frequency. It was postulated that such frequency estimates are meaningful only if they are predicated upon optimizing behavior. In that case the desired frequency estimate must be the locus of the maximum of total preference as a function of eating frequency. Indeed, when school lunches were planned on the basis of a quadratic

approximation of these preference-frequency functions, double blind experiments showed a significant increase in participation and food consumption [Balintfy, Rumpt, Sinha 2].

The purpose of this article is to show that the preference-time and preference-frequency relations are equivalent properties of repetitive consumption and the functional forms of these relations follow a well-defined rule of correspondence. Further analysis of these equivalence relations appears to explain variety seeking behavior in terms of mathematical optimization models.

The equivalence relations identified in this paper provide the theoretical foundation for selecting the appropriate utility functions and coefficient estimation procedures in related surveys. These utility functions are very important in optimizing menu planning and scheduling decisions by mathematical programming techniques for a variety of institutional food service systems. Experimental evidence indicates that preference maximized menus do increase satisfaction and/or decrease food cost significantly in practice [3].

DEFINITIONS:

a. $f(t)$ is the preference-time function where t is the time interval at which a unit portion of a food is consumed.

b. $x = T/t$ is the frequency of consumption over a time horizon T.

c. $g(x)$ is the preference-frequency function representing the total preference obtained during a time period T by consuming a unit portion of an item x times repetitively at $t = T/x$ times apart during that period.

d. $g(x) = xf(t) = xf(T/x),\ T > 0.$

Graphs of the functions $f(t)$ and $g(x)$ are illustrated below.

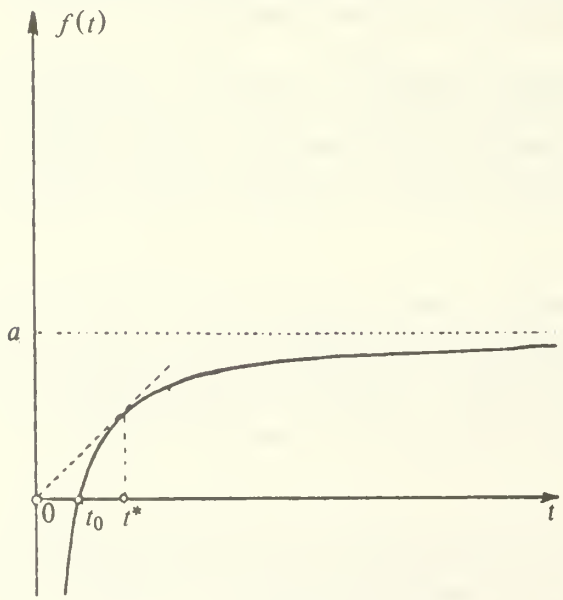


Figure 1. Preference-Time Function

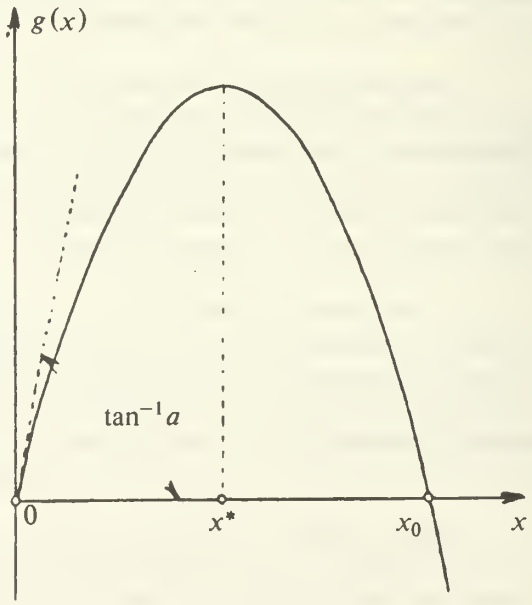


FIGURE 2. Preference-Frequency Function

PROPOSITION 1: If $f(t)$ is a twice differentiable continuous function defined for $t \geqslant 0$, strictly concave, monotonically increasing, bounded from above and has a positive root, i.e., $f(t_0) = 0$ for $t_0 > 0$, then

(i) $\lim_{t \to \infty} tf'(t) = 0$

(ii) The function $g(x) = xf(T/x)$ with $T > 0$ is strictly concave for $x \geqslant 0$

(iii) The function $g(x)$ obtains its maximum at a unique finite point x^* which is a solution of the equation $f(T/x) = (T/x)f'(T/x)$.

(See Figures 1 and 2.)

PROOF:

(i) Since $f'(t) \geqslant 0$, it is clear that $t \cdot f'(t) \geqslant 0$. Let us assume that there exist $\epsilon > 0$ and a sequence $t_n \to \infty$ for which $t_n f'(t_n) \geqslant \epsilon$. We may assume for a subsequence that $t_{n+1}/t_n > 2$, i.e., $t_n/2 > t_{n-1}$. Then $f'(t_n) \geqslant \epsilon/t_n$ and since $f'(t)$ is a decreasing function, it is implied that $f'(t) \geqslant \epsilon/t_n$ for $t \leqslant t_n$. Consequently, $f(t_n) - f(t_n/2) \geqslant (\epsilon/t_n)\,(t_n/2) = \epsilon/2$. Since the $(t_n/2,\ t_n)$ intervals are not overlapping, $f(t_n) \geqslant \sum_{i=1}^{n} \{f(t_i) - f(t_i/2)\} + f(t_1/2) \geqslant n(\epsilon/2) + f(t_1/2)$ which contradicts the assumption that $f(t)$ is bounded.

(ii) We need only to prove that the second derivative $g''(x)$ is negative for $x \geqslant 0$.

Let

$$g'(x) = f(T/x) - (T/x)\, f'(T/x) \tag{1}$$

Then

$$g''(x) = (T^2/x^3) f''(T/x).$$

Since $f(t)$ is strictly concave, $f''(T/x) < 0$. The nonnegativity assumption on x implies that $(T^2/x^3) > 0$, which preserves the inequality, i.e., $g''(x) < 0$ is proved.

(iii) Since $g(x)$ is strictly concave, its unique maximum occurs where its first derivative vanishes, i.e., $g'(x) = 0$. This implies by (ii) that: $f(T/x) = (T/x)f'(T/x)$, or by substituting $t = (T/x)$,

$$f(t) = tf'(t). \tag{2}$$

To prove that (2) has a positive solution, $t^* = (T/x^*) > 0$, implying that the corresponding x^* is finite, let us consider the expression $h(t) = f(t) - tf'(t)$. The function $h(t)$ is < 0 for $t \leqslant t_0$, since $f(t_0) = 0$, and positive for $t > t^*$, since $t \cdot f'(t) \to 0$. By continuity, the equation $h(t) = 0$ must have a solution for $t = t^* > t_0$. At that point, t^*, $g(x)$ attains its maximum, since the second derivative $g''(x)$ is strictly negative at $x^* = T/t^*$:

$$g''(x^*) = (T^2/x^{*3}) f''(T/x^*) = (t^{*3}/T) f''(t^*) < 0.$$

COROLLARY 1: The tangent of $g(x)$ at $x = 0$ is equal to the upper bound of $f(t)$, i.e., $g'(0) = \lim_{t\to\infty} f(t) = a$. Taking the limit in (1) as $x \to 0$, and taking into account (i)

$$g'(0) = \lim_{x\to 0} \{f(T/x) - (T/x)f'(T/x)\} = a$$

since

$$\lim_{x\to 0} f(T/x) = \lim_{t\to\infty} f(t) = a \text{ and } \lim_{x\to 0} (T/x)f'(T/x) = \lim_{t\to\infty} tf'(t) = 0.$$

COROLLARY 2: The value of x at which $g(x)$ obtains its maximum is given by $x^* = (T/t^*)$ where t^* is the abscissa of the point where the line from the origin is tangent to $f(t)$. By (2), $f(t)/t = f'(t)$ at $t = t^*$, i.e., the mean of the function is equal to its derivative. This is also the point where $f(t)/t$ reaches its maximum and the cord from the origin becomes the tangent.

PROPOSITION 2: If $g(x)$ is a continuous twice differentiable concave function, with a zero root, i.e., $g(0) = 0$, and bounded from above, then $f(t)$, defined as $f(t) = (t/T)\, g(T/t)$ where T is a positive constant, $t > 0$, has the following properties: (i) it is strictly concave, (ii) it is monotonically increasing, (iii) it is bounded from above, and furthermore, (iv) if $g(x)$ has a maximum at a positive finite point, then $f(t)$ has a positive root.

PROOF:

(i) In order to show the concavity of $f(t)$, it is sufficient to prove that $f''(t) < 0$ for $t \geqslant 0$

$$f'(t) = (1/T)g(T/t) - (1/t)g'(T/t)$$

$$f''(t) = (T/t^3)\, g''(T/t).$$

Strict concavity of $g(x)$ implies $g''(x) < 0$ for every x. For $t \geqslant 0$, $(T/t^3) > 0$, preserving the inequality. So $f''(t) < 0$ or $f(t)$ is a strictly concave function.

(ii) In order to show that $f(t)$ is monotonically increasing, it is sufficient to prove that $f'(t) > 0$ for $t \geqslant 0$, or by (3):

$$g(T/t) > (T/t)g'(T/t). \tag{4}$$

Using the mean value theorem for the points (T/t) and 0, we get:

$$g(T/t) - g(0) = (T/t)g'(T/\tau)$$

where

$$0 < (T/\tau) < (T/t). \tag{5}$$

Since it is given that $g(0) = 0$, (4) becomes: $(T/t)g'(T/\tau) > (T/t)g'(T/t)$, or since $(T/t) > 0$, then $g'(T/\tau) > g'(T/t)$ which is always true for a strictly concave function because of (5).

(iii) Since $g(x)$ is bounded from above, and $g(0) = 0$, there always exists a large number, say a, such that $g(x) \leqslant xa$ for $x \geqslant 0$ substituting $x = (T/t)$ $(t \geqslant 0)$, we get $g(T/t) \leqslant (T/t)a$ or $f(t) = (t/T)g(T/t) \leqslant a$ i.e., $f(t)$ is bounded from above.

(iv) For the last part, we have only to prove that the equation $f(t) = 0$, or by definition the equation $(t/T)g(T/t) = 0$ has a solution for $t > 0$. That can be done only if:

$$g(T/t) = 0 \text{ for } t > 0 \tag{6}$$

However, since $g(x)$ starts from 0, i.e., $g(0) = 0$, and obtains its maximum at a positive finite point, being always strictly concave for $x \geqslant 0$, there will exist a positive $x = x_0 > x^*$ for which it becomes 0 again. In other words, there exists a $t_0 > 0$ such that $(T/t_0) = x_0 > 0$, for which (6) holds.

Applying the above properties to well-known results concerning the maximization of concave functions, the following two corollaries are stated without proof:

COROLLARY 3: If a time horizon T is partitioned into t_i time intervals, such that $\sum_{i=1}^{m} t_i = T$, and the preference-time function $f(t_i)$ for a given item is a concave twice differentiable function, then optimal repetitive consumption strategy over time implies that all the time intervals must be equal, i.e., $t_i = t_j$ for all i and j. This is to say that at optimum the marginal preferences with respect to time must be equal, and that is equivalent with the proposition that at optimum the time intervals of consuming the same food are equal.

COROLLARY 4: Given concave unimodal preference-frequency $g_j(x_j)$ for different food items $j = 1, 2, \ldots, n$, and positive quantity K such that $\sum_{j=1}^{m} x_j = K$, then the consumption frequencies of the foods, x_j, are optimal if the marginal preference with respect to the eating frequency is equal for each item, provided $x_j > 0$. This solution is feasible only if $g_j'(x_j) = \lambda$ has a positive root for all j, which is not necessarily assured.

COROLLARY 5: Optimal consumption frequencies that are all positive and all integers always exist if K is integer.

PROOF: The corollary implies that the following nonlinear integer programming problem always has solution:

P1 maximize $\sum_{j=1}^{n} g_j(x_j)$

subject to $\sum_{j=1}^{n} x_j = K$

$x_j \geqslant 0$

x_j, K are integers.

Let x_j be expressed by the auxiliary variable y_{ij} where

$$x_j = \sum_{i=0}^{m} i y_{ij}, \qquad j = 1, 2, \ldots, n$$

subject to $\sum_{i=0}^{m} y_{ij} = 1$

$y_{ij} \geqslant 0$

assuming that x_j is bounded by the intervals $0 \leqslant x_j \leqslant m$ for all j.

Similarly

$$g_j(x_j) = \sum_{i=0}^{m} g_j(i) y_{ij} .$$

Then we have the following piecewise linear programming representation of P1:

P2 maximize $\sum_{j=1}^{n} \sum_{i=0}^{m} g_j(i) y_{ij}$

subject to $\sum_{j=1}^{n} \sum_{i=0}^{m} i y_{ij} = K$

$$\sum_{i=0}^{m} y_{ij} = 1 \qquad j = 1, 2, \ldots n.$$

$$y_{ij} \geqslant 0.$$

The above linear programming problem has $n + 1$ rows and consequently an $(n + 1) \times (n + 1)$ basis. This means that for $(n - 1)$ variables, $n - 1$ convexity constraints can be satisfied if, and only if, $y_{ij} = 1$ yielding $n - 1$ integer frequencies. The remaining two variables are, however, on adjacent grid points of the same item due to the concavity of $g_j(x_j)$ and according to the simplex rule of optimization. Let these points be i and $i + 1$. Then the subproblem:

$$i y_{ij} + (i + 1) y_{i+1,j} + N = K$$

where $N < K$, and both N and K are integers, can have positive solution if, and only if, either y_{ij} or $y_{i+1,j}$ is zero. Consequently, P2 must have a degenerate solution which assures that all the frequencies of solving P1 are nonnegative integers. Thus, the optimum consumption frequencies for P1 can always be determined by linear programming.

CONCLUSIONS

The empirical evidence concerning the shape of preference-time and preference-frequency functions was utilized to derive results to show that these functions are describing equivalent mathematical properties. Consequently, one of the functions can always be derived from the other.

This finding is important for the consistent estimation of the parameters of the functions chosen for the analytical representation of the phenomena under study. If, in the simplest case, the perference-frequency function is approximated by a quadratic function, then the preference-time function will be hyperbolic. In this case, it is sufficient to estimate the upper bound of $f(t)$ and the locus of the maximum of $g(x)$, i.e., one parameter from each function. For more complex functions the precision of curve fitting can be improved by imposing the equivalent properties of both functions on the data points used.

Another important result is the theoretical foundation this investigation yields for the proper formulation of objective functions for planning human diets. The augmentation of P1 or P2 with nutritional and other desired constraints leads to the formal statement of preference maximizing menu planning problems further described in Balintfy and Sinha [4]. The solution of these problems has a great importance in institutional menu planning and in individual diet planning.

In a more general sense, the existence of preference-time functions explains why repetitive consumption may yield optimal consumption frequencies. Vice versa, whenever variety seeking consumption behavior occurs, the presence of preference-time functions can be rightly assumed.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of Professor L.A. Rubel of the University of Illinois for contributing the proof of Proposition 1, part (i), and Professor P. Sinha at Rutgers University for suggesting the proof of Corollary 5.

REFERENCES

[1] Balintfy, J.L., W.J. Duffy and P. Sinha, "Modeling Food Preferences Over Time," Operations Research, *22*, 5, 711-727 (1974).

[2] Balintfy, J.L., D.L. Rumpt and P. Sinha, "The Effect of Preference Maximized Menus on the Consumption of School Lunches," School Food Service Research Review, *4*,1, 48-53 (1980).

[3] Balintfy, J.L., K. Jarrett, F. Paige and P. Sinha, "Comparison of Type A Constrained and RDA Constrained School Lunch Planning Computer Models," School Food Service Research Review, *4*,1, 54-62 (1980).

[4] Balintfy, J.L. and P. Sinha, "A Nonlinear Programming Approach to Preference Maximized Menu Plans," in *Computers and Mathematical Programming*, W.W. White, Editor, NBS Special Publication 502, U.S. Dept. of Commerce, National Bureau of Standards, 1978, 97-105.

A SAMPLING PROCEDURE AND PUBLIC POLICY

G. J. Lieberman*
I. Olkin
and
F. Riddle

Stanford University
Stanford, California

ABSTRACT

This paper describes the background of the Office of Management Budget Circular A-21, "Principles for Determining Costs Applicable to Grants, Contracts, and Other Agreements with Educational Institutions," that describes the requirement for effort reporting. A sampling procedure is proposed as an alternative to 100% reporting.

1. INTRODUCTION

Following the March, 1979 publication of the revised Office of Management Budget Circular A-21 "Principles for Determining Costs Applicable to Grants, Contracts, and Other Agreements with Educational Institutions,"† which promulgates the cost principles to be used in the determination of cost charged to government sponsored agreements, there was an undercurrent of faculty dissatisfaction with the revised rules. Of major issue was, and still is, the requirements relating to reporting of faculty effort charged directly or indirectly to government agreements.

The two acceptable methods for effort reporting contained in the revised A-21 are contained in section J.6 of that document and are called Monitored Workload and Personal Activity Reporting. The Monitored Workload method can be used for faculty and professionals only, whereas the Personal Activity Reporting method can be used for all personnel.

The major objection raised by the faculty centered on the requirement (under either system) to account for 100% of their effort applicable to each sponsored agreement, each indirect cost category, and each major function of the University. This level of accountability was considered burdensome, arbitrary, and to some extent, useless work because of the interrelated nature of the research, instruction and other activities of the faculty. The results of such effort reports, though fulfillng the documentation requirements established by the regulation, were known to be arbitrary and therefore inaccurate, and with the added detail terribly expensive for institutions to implement.

That the assessment of direct and indirect costs is, at times, highly subjective and approximate was also recognized by the government. Section J.6.b of OMBA-21 states:

*This work was supported, in part, under Contract N00014-75-C-0561 with the Office of Naval Research.
†These cost principles were revised again in July 1982.

> "... it is recognized that, because of the nature of the work involved in academic institutions, the various and often interrelated activities of professorial and professional employees frequently cannot be measured with a high degree of precision, that reliance must be placed on reasonably accurate approximations and that acceptance of a degree of tolerance in measurement is appropriate."

Because of dissatisfaction with the new effort reporting guidelines, the major research institutions formed a task force under the auspices of the Association of American Universities (AAU) to develop alternative methods of gathering faculty effort without the burden necessitated by the new regulations. As a consequence of their deliberations, it was suggested that a method of statistical sampling be used to determine the indirect effort performed by the faculty. The results of the sample could then be used in determining indirect cost rates.

It was also noted that too often the approximations made in effort reporting resulted in very poor data. To improve the data, the faculty would have to be apprised of what specific costs are to be labeled a direct or indirect cost. By cutting the scope of reporting, a sampling procedure would permit increased dialogue between the administration and faculty thereby enabling the faculty to have a better understanding of the process.

The idea of statistical sampling was explored between Stanford University and the Office of Management and Budget. The OMB felt the idea had possibilities and suggested that Stanford submit a formal proposal. The authors carried out some preliminary work to describe how such a system would work, and a plan was presented to OMB and a panel of Federal statisticians. After their review, agreement was reached to proceed with a more detailed study to include the process, the sample sizes to reach a 90% confidence level with a probable 5% error and the statistical formulas to produce the desired data.

The plan was completed in January, 1981 and submitted and accepted by OMB which had chosen ten universities who had expressed interest in using the statistical method as test of the proposal. Unfortunately, of those institutions who submitted their proposals based on the statistical model, the implementing agency wanted the test made in parallel with a standard J.6 method so that the results could be verified. In essence, this stringent requirement virtually terminated further consideration by the institutions.

However, on January 7, 1982, after several months of added effort by the AAU and other higher education associations, a proposal for revision of section J.6. of A-21 was placed in the *Federal Register.* Included in the proposed revision is paragraph J.6.b(2) (c), which reads:

> "(c) The payroll distribution system will allow confirmation of activity allocable to each sponsored agreement and each category of activity needed to identify indirect costs and the functions to which they are applicable. The activities chargeable to indirect cost categories or the major functions of the institution for employees whose salaries must be apportioned (see J.6.b(1) (b) above) if not initially identified as separate categories, may be subsequently distributed by any reasonable method mutually agreed to, including, but not limited to, suitably conducted surveys, *statistical sampling* (emphasis added) procedures, or the application of negotiated fixed rates."

It is assumed that the public comment will be overwhelmingly favorable to the proposed revision, which makes the publication of the method developed at Stanford by the authors

timely. Use of the method would allow confirmation of salaries directly charged to sponsored agreements by confirming notations on the monthly expenditure reports and less costly (and less disruptive for the faculty), gathering of data necessary for indirect costing purposes. What follows is the proposal that was submitted to OMB as a statistical sampling model.

2. THE SAMPLING PLAN

The proposed sampling plan is designed for a particular university for a particular purpose. However, the method can be extended to other organizations and for other purposes.

The sample is designed to determine the amount of effort expended by the faculty in indirect activities such as departmental research, university and student service administration. For example, if the indirect activity departmental administration is of interest, then we need to estimate the total cost, T, expended by the faculty engaged in that activity

$$T = f_1 s_1 + \cdots f_N s_N,$$

where f_j is the fraction of time that the jth faculty members devote to departmental administration, s_j is the academic year salary of the jth faculty member and N is the number of relevant faculty. There will be similar totals for university administration, research administration, etc.

Because faculty differ with respect to their activities, we use a stratified sample thereby reducing the sample variance of our estimates. Two methods of stratification appear natural in a university context, namely, rank and disciplines.

For stratification by rank four strata may be appropriate: professors, associate professors, assistant professors, department chairs and faculty with major administrative duties. Stratification by discipline might include the Schools of Medicine, Engineering, and Education, the Physical Sciences, Social Sciences, the Humanities, etc. In making these divisions it is important to emphasize that within each stratum the faculty should be relatively homogeneous.

Definition of the Universe

The universe is assumed to be all faculty who conduct sponsored activities and have other duties funded from nongovernment sources, including the departmental operating budget. Also included would be faculty totally funded from the operating budget or a combination of operating budget and other nongovernment sources. A faculty person funded 100% on sponsored agreements or patient care would be excluded from the universe. Since the results of the sample need to be generalized to the universe it represents, the sample must reflect the elements necessary to fulfill the objective of the universe.

The Strata

At Stanford University strata were chosen using a combination of both the rank and discipline criteria. These were:

(1) Medical School department chairs.

(2) Other department chairs and laboratory directors.

(3) Full professors in the sciences.

(4) Full professors in other fields.

(5) Associate professors.

(6) Assistant professors.

3. DETAILS OF THE SAMPLING PLAN

3.1 Preliminaries

For each stratum, we have available the total number, N_i, of faculty members, and the average, $\bar{Y}_i$, of the product of academic year salary and proportion of time devoted to departmental administration. Consequently, we can also determine the sample variance, S_i^2, of the Y-values. That is

$$S_i^2 = \Sigma(Y - \bar{Y}_i)^2/(N_i - 1).$$

Actually, much of the above data is unknown beforehand. However, some preliminary estimates are essential if we are to sample with a specified degree of accuracy. One procedure is to obtain estimates from past data, or from a sample of past data. Indeed, an examination of past data may also lead to improved stratification. At Stanford an examination of 1979-1980 data was used to estimate the needed parameters and also to establish the use of the strata.

3.2 Choice of Sample Sizes

The design depends on choosing an overall sample size, n, and an allocation into the various strata. The total sample size depends on the confidence level, c, and a given relative error, e. The higher the required confidence, the larger the required sample size; the smaller the relative error, or equivalently, the greater the precision, the larger the required sample size. The sample size also depends on the variances in each stratum, but not in a simple way.

The determination of n is given by

$$n \cong \frac{(\Sigma N_i S_i)^2}{W^2 + \Sigma N_i S_i^2} = \frac{T^2}{W^2 + V}, \tag{1}$$

where $T = \Sigma N_i S_i$, $V = \Sigma N_i S_i^2$ and W depends on the confidence level, c, and the relative error e:

$$W = e\,(\text{Total})/c^*, \tag{2}$$

where the total is the estimated true total cost of administration, $c^* = \Phi^{-1}((1 + c)/2)$, and Φ is the standard normal cumulative function.

Once the sample size is chosen, the allocation n_i to stratum i is made in the proportion

$$n_i = \left(\frac{N_i S_i}{\Sigma N_j S_j}\right) n, \qquad n_i \leqslant N_i. \tag{3}$$

3.3 An Example

The following data with five strata are fictitious.

Stratum	Number N_i in stratum	Variance S_i^2 in each stratum (in millions)	Stratum standard deviation S_i (in thousands)	N_iS_i (thousands)	$\frac{N_iS_i}{\Sigma N_jS_j}$	$N_iS_i^2$ (in millions)
A	B	C	D	E	F	G
1	20	100.00	10.0	200	.110	2000
2	50	4.00	2.0	100	.055	200
3	200	9.00	3.0	600	.329	1800
4	250	6.25	2.5	625	.342	1562.5
5	300	1.00	1.0	300	.164	3.0
				T=1,825	1.000	V=5,862.5

If we wish to estimate the total amount spent on departmental administration to within 5% of the true total with 90% confidence, and if we estimate the true total to be more than $7 million, then from (2)

$$W = \frac{(.05)\ (7,000,000)}{1.645} = 212,766,$$

where $c^* = 1.645$ is obtained from the tables of normal distribution. From (1)

$$n \geqslant \frac{(1,825,000)^2}{(212,766)^2 + (5,862,500,000)} \cong 65.1,$$

which we round upward to $n = 66$.

The allocation to strata is now obtained from the proportions in column F. Thus, stratum one will have a sample size 66(.11) = 7.3. The results for the five strata are

7.3, 3.6, 21.7, 22.6, 10.8.

Since these values are not integers, it will be safest to round up to

8, 4, 22, 23, 11,

which yields a total sample size of 68.

Because Stanford is on a quarter system, the sample sizes within strata are divided into thirds, representing the three quarters. To avoid unequal numbers per quarter, and to simplify the procedure, we round upwards so that the samples are divisible by 3, namely,

9, 6, 24, 24, 12.

Thus, for example, we would sample four per quarter in the fifth stratum. The total sample size would now be 75.

Remark

An alternative procedure, if a sample is not divisible by 3, is to randomize the remainder. For example, in the fifth stratum, take three faculty per quarter and assign the remaining two at random to two quarters.

4. DERIVATIONS AND TECHNICAL RESULTS

We now provide some of the derivations and rationale for the formulas. Formulas that are standard are not derived.

Let y_{ijk} be the y value of the jth faculty member in stratum i for quarter (semester) k. [Since Stanford is on a quarter system, we use three quarters in our discussion. However the procedures are easily modified for use at a university with the semester system.]

A sample of size n is chosen and then partitioned in accordance with the allocation

$$n_i = \frac{N_i S_i}{T} n, \qquad n_i \leqslant N_i,$$

in each stratum, where we assume that n_i is rounded upward to be divisible by 3 (or 2). We then measure the y-value of 1/3 (or 1/2) of the sample in each stratum each quarter (or semester).

A notation is needed to denote measurements taken in each of the quarters, and we use $\dot{y}_i$, and $\ddot{y}_i$ and $\dddot{y}_i$ to denote the sums of y-values in quarters 1, 2, 3, respectively. That is

$$\dot{y}_i = \sum_{j=1}^{\bar{n}_i} y_{ij1},$$

$$\ddot{y}_i = \sum_{j=\bar{n}_i+1}^{2\bar{n}_i} y_{ij2}$$

$$\dddot{y}_i = \sum_{j=2\bar{n}_i+1}^{n_i} y_{ij3},$$

where $\bar{n}_i = n_i/3$.

The sampling procedure has two-stages: In the first stage we sample the faculty within each stratum, and in the second stage, sample by quarters (semesters). The quantity to be estimated is

$$Y \equiv \sum_{i=1}^{g} \sum_{j=1}^{N_i} \sum_{k=1}^{3} Y_{ijk}/3, \tag{3}$$

where g is the number of strata. The estimate used is

$$\hat{Y} = \sum_i \frac{N_i}{n_i} (\dot{y}_i + \ddot{y}_i + \dddot{y}_i), \tag{4}$$

which has variance

$$V(\hat{Y}) = V(E(\hat{Y}|\text{sample})) + E(V(\hat{Y}|\text{sample})). \tag{5}$$

Presumably, $V(E(\hat{Y}|\text{sample}))$ is large compared to $E(V(\hat{Y}|\text{sample}))$.

Consider the population with values

$$Y_{ij}^* = \frac{1}{3}(y_{ij1} + y_{ij2} + y_{ij3}).$$

The total Y^* is just Y, and our sampling scheme reduces to a single-stage, stratified simple random sample. Also,

$$\hat{Y}^* = \sum_i \frac{N_i}{n_i^*} \sum_{j=1}^{n_i} y_{ij}^*,$$

which has variance

$$\sum_i \frac{N_i^2 (S_i^*)^2}{n_i^*} \left(1 - \frac{n_i^*}{N_i}\right) = V(\hat{Y}^*).$$

With optimal allocation, $n_i^* = n(N_i S_i^*/T^*)$, where $T^* = \Sigma N_\alpha S_\alpha^*$. Then

$$V(\hat{Y}^*) = \sum_i \frac{N_i^2 S_i^{*2}}{n \dfrac{N_i S_i^*}{T^*}} \left(1 - \frac{\dfrac{N_i S_i^*}{T^*} n}{N_i}\right) = \sum_i \frac{T^* N_i S_i^*}{n} \left(1 - n \frac{S_i^*}{T^*}\right)$$

$$= \sum_i N_i S_i^* \left(\frac{T^*}{n} - S_i^*\right) = \frac{T^{*2}}{n} - V^*,$$

where $V^* = \Sigma N_i S_i^{*2}$. Now, $S_i^* \leqslant S_i$, so that

$$V(\hat{Y}^*) \leqslant \frac{T^2}{n} - V.$$

Also, because the expectation is linear,

$$V(E(Y|\text{sample})) = V\left(\sum_i \frac{N_i}{n_i} \sum_{j=1}^{n_i} y_{ij}^*\right) = V(\hat{Y}^*).$$

Consequently, $V(\hat{Y}) \geqslant V(\hat{Y}^*)$, and

$$V(\hat{Y}) \doteq V(Y^*) \leqslant \frac{T^2}{n} - V,$$

which suggests the approximation

$$\frac{T^2}{n} - V \doteq V(\hat{Y}).$$

For relative error e with confidence coefficient c, we must have

$$P\{Y(1-e) \leqslant Z \leqslant Y(1+e)\} \geqslant c,$$

where Z is normally distributed with mean Y and variance $V(Y) \doteq (T^2/n) - V$. This requires that the condition

$$2\Phi(W) \equiv 2\Phi\left(eY/\sqrt{V(Y)}\right) - 1 \geqslant c \tag{7}$$

be satisfied. But (7) is equivalent to

$$eY/\sqrt{V(Y)} \geqslant \phi^{-1}\left(\frac{1+c}{2}\right),$$

where Φ is the standard normal cumulative distribution. Consequently,

$$V(\hat{Y}) \doteq \frac{T^2}{n} - V \leqslant W^2,$$

from which we obtain

$$n \geqslant \frac{T^2}{W^2 + V}. \tag{8}$$

This formula is slightly optimistic because it allows for fractional allocation to strata and because $V(\hat{Y})$ was approximated. Since T and V are not known exactly, further caution would be used in picking n.

ACKNOWLEDGMENT

The authors wish to acknowledge a number of conversations with colleagues concerning the concept of statistical sampling in this context. In particular, we are exceedingly grateful to Peter Bacchetti who helped immeasurably with the details and computations of the sampling plan.

BIBLIOGRAPHY

[1] Cochran, W.G., *Sampling Techniques* (Third Edition), (John Wiley and Sons, New York, 1977).

[2] Office of Management and Budget. Circular A-21 (Revised), "Principles for Determining Costs Applicable to Grants, Contracts and Other Agreements with Educational Institutions," (July 23, 1982).

THE LOGARITHMIC POISSON GAMMA DISTRIBUTION: A MODEL FOR LEADTIME DEMAND

Steven Nahmias

Department of Decision and Information Sciences
University of Santa Clara
Santa Clara, California

W. Steven Demmy

Department of Management
Wright State University
Dayton, Ohio

ABSTRACT

This paper considers an inventory system in which demand occurrences arise according to a stationary Poisson process, demand sizes at each occurrence follow a logarithmic distribution, and leadtimes are random variables with the gamma distribution. Both the exact and approximate distribution for leadtime demand are derived and computations are performed which compare the approximation to the exact distribution. The results have application to both repairable and consumable item inventory systems.

1. INTRODUCTION

Many of the inventory models which are used in practice reply upon knowing the probability distribution of demand over a leadtime. The common assumption is that this distribution is normal. However, in certain circumstances, the normality assumption may in inappropriate. The purpose of this paper is to derive the exact distribution of leadtime demand under the following assumptions: customer requisitions occur according to a stationary Poisson process, requisition sizes follow a logarithmic distribution and leadtime is a random variable with the gamma distribution. In addition to deriving the exact distribution of leadtime demand, we compare our results to actual operational data and discuss a variety of approximations.

A number of researchers have considered the problem of determining inventory operating policies when requisition size exceeds one. For example, Hausman [6] extends Hadley and Whitin's [5] heuristic while Archibald and Silver [1] derive optimal (s,S) policies. These studies differ from ours in two ways. First, in every case leadtime was assumed to be deterministic. Second, they focus primarily on describing optimal and suboptimal ordering policies. Our interest is in a detailed examination of the distribution of demand over leadtime.

2. THE LOGARITHMIC DISTRIBUTION

The logarithmic (or log series) distribution was originally derived by Fisher, Corbet and Williams [4] and has been discussed by Sherbrooke [11] in connection with inventory problems. It can be derived as a limiting case of the negative binomial distribution and has the form

$$f(x) = \frac{\theta^x}{-x \ln(1-\theta)} \quad \text{for } x = 1, 2, \ldots \quad \text{for } 0 < \theta < 1. \tag{1}$$

A maximum likelihood estimator for θ is obtained as follows. The likelihood function, $L(\theta)$, is

$$L(\theta) = \theta^{n\bar{x}}[-ln(1-\theta)]^{-n} \tag{2}$$

$$L'(\theta) = n\bar{x}\,\theta^{n\bar{x}-1}[-\ln(1-\theta)]^{-n} + \theta^{n\bar{x}}(-n)[-ln(1-\theta)]^{-n-1} \cdot \frac{1}{1-\theta}.$$

The maximum likelihood estimator, $\hat{\theta}$, solves $L'(\hat{\theta}) = 0$ or

$$n(\hat{\theta})^{n\bar{x}-1}[-\ln(1-\hat{\theta})]^{-n-1}\{\bar{x}(-ln(1-\hat{\theta}))-\hat{\theta}/(1-\hat{\theta})\} = 0$$

which gives

$$\bar{x} = \frac{\hat{\theta}}{-(1-\hat{\theta})\ln(1-\hat{\theta})} \tag{3}$$

where $\bar{x}$ is the observed sample mean. (This estimator could have also been derived by the method of moments as suggested by Chakravarti, Laha and Roy [3].) Note that this is a transcendental equation in $\hat{\theta}$ which can be solved efficiently by interval bisection since the right hand side is an increasing function of θ. Alternately, one can use a recursive scheme as follows: let $\omega = \theta/1-\theta)$ then (3) is equivalent to $\bar{x}\ln(1+\omega) = \omega$. This suggests the recursion $\omega_{n+1} = \bar{x}\ln(1+\omega n)$. As an example, the observed sample mean of the data represented in Table 1 is 3.94. Starting with $\omega_0 = .5$ we obtain

$$\omega_1 = 1.5975$$
$$\omega_2 = 3.76$$
$$\omega_3 = 9.08328$$
$$\vdots$$
$$\omega_7 = 9.11864$$

which results in $\hat{\theta} = \omega_7/(1+\omega_7) = .9012$. Using this value in (1) we can compare the observed and theoretical frequencies for a typical item. These results, which are reported in Table 1 show the close agreement between the observed and predicted cumulative distribution functions for this item.

TABLE 1 — *Comparison of Observed Frequencies and Those Predicted by Logarithmic Distribution for a Typical EOQ Type Item*

x	Number of Observations	Observed Frequency	Theoretical Frequency	Observed Cumulative	Theoretical Cumulative
1	93	.4247	.3896	.4247	.3896
2	31	.1416	.1755	.5663	.5651
3	13	.10594	.1054	.6257	.6705
4	15	.0685	.0712	.6942	.7417
5	10	.0457	.0514	.7399	.7931
6	15	.0685	.0386	.8084	.8317
7	8	.0365	.0298	.8449	.8615
8	8	.0365	0.235	.8814	.8850
9	3	.0137	.0188	.8951	.9031
10	4	.0183	.0152	.9134	.9190
11	7	.0320	.0125	.9454	.9315
12	3	.0137	.0103	.9591	.9418
13	0	.0000	.0086	.9591	.9504
14	1	.0046	.0072	.9637	.9576
15	2	.0091	.0061	.9728	.9637
16	1	.0046	.0052	.9774	.9689
17	0	.0000	.0043	.9774	.9732
18	1	.0046	.0037	.9820	.9769
19	0	.0000	.0031	.9820	.9800
20	2	.0091	.0027	.9911	.9827
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
25	2	.0091	.0008	1.0000	.9915

3. THE LPG DISTRIBUTION

Let us now assume that requisitions are generated by a Poisson process and the requisition size has a logarithmic distribution. (That is, the demand process is a compound Poisson process with logarithmic compounding distribution.) It is well known that (see, for example, Quenouille [10]) that the total number of units demanded in any fixed time t, say $Z(t)$, has the negative binomial distribution. In particular, we obtain

$$P\{Z(t) = x\} = \frac{\Gamma(ct + x)}{x!\,\Gamma(ct)}(1 - \theta)^{ct}\theta^{x} \quad \text{for } x = 0, 1\ 2, \ldots \tag{4}$$

where $c = -\lambda/ln(1 - \theta)$ and λ is the requisition arrival rate.

Boswell and Patil [2] give fifteen different derivations of the negative binomial distribution, thus accounting for its power in describing many common phenomena.

Now let us assume that the procurement leadtime, τ, is a continuous nonnegative random variable with probability density $g(\tau)$. In general, the number of units demanded during a random leadtime τ is a random variable with probability function $h(x)$ given by

$$h(x) = \int_0^\infty f(x|\tau)\ g(\tau)\,d\tau \tag{5}$$

where $f(x/\tau)$ is the probability function of the number of units demanded in a time τ. Under our assumptions, $f(x/\tau)$ has the negative binomial distribution. Since $c\tau$ is, in general, not an integer, we use the gamma function representation for the factorials, so that

$$h(x) = \frac{\theta^x}{x!} \int_0^\infty \frac{\Gamma(c\tau + x)}{\Gamma(c\tau)} (1-\theta)^{c\tau} g(\tau) d\tau. \tag{6}$$

Using the fact that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, we have

$$\frac{\Gamma(c\tau + x)}{\Gamma(c\tau)} = \prod_{j=0}^{x-1} (c\tau + j) = \sum_{k=1}^{x} (c)^k S_{xk}, \tag{7}$$

where the coefficients S_{xk} are known as Stirling numbers of the first kind and can be computed from the recursion

$$S_{xk} = S_{x-1,k-1} + (x-1)\, S_{x-1,k}, \tag{8}$$

for $k = 1, 2, \ldots, x$ and $x = 1, 2, \ldots,$

with $S_{x0} = 0$ for all x.

Furthermore, from the definition of c,

$$(1-\theta)^{c\tau} = \exp\{c\tau \ln(1-\theta)\} = e^{-\lambda\tau}, \tag{9}$$

so that we may now write

$$h(x) = \frac{\theta^x}{x!} \sum_{k=1}^{x} c^k S_{xk} \int_0^\infty \tau^k e^{-\lambda\tau} g(\tau) d\tau. \tag{10}$$

We now specialize to the case where $g(\tau)$ has the gamma distribution with parameters α and β so that

$$g(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} \text{ for } \tau > 0. \tag{11}$$

Since leadtimes must be nonnegative, the gamma distribution should provide sufficient flexibility to model leadtime variability in many operating environments.

Using the fact that

$$\int_0^\infty \tau^{k+\alpha-1} e^{-(\lambda+\beta)\tau} = \Gamma(k+\alpha)/(\lambda+\beta)^{k+\alpha} \tag{12}$$

and that, as above,

$$\frac{\Gamma(k+\alpha)}{\Gamma(\alpha)} = \sum_{j=1}^{k} \alpha^j S_{kj}, \tag{13}$$

we obtain the following as the probability function for the number of units demanded in a leadtime:

$$h(x) = \begin{cases} (\beta/(\lambda+\beta))^\alpha & \text{for } x = 0 \\ \left(\dfrac{\beta}{\lambda+\beta}\right)^\alpha \dfrac{\theta^x}{x!} \displaystyle\sum_{x=1}^{x} \left(\frac{c}{\lambda+\beta}\right)^k S_{xk} \sum_{j=1}^{k} \alpha^j S_{kj} & \text{for } x = 1, 2, 3, \ldots \end{cases} \tag{14}$$

We call this the LPG distribution (for Logarithmic-Poisson-Gamma). Its four parameters are α, β, θ and λ (θ and λ determine c). An example of the LPG distribution is presented in Table 2 for $\alpha = 1$, $\beta = 1$, $\theta = .8$ and $\lambda = 1$.

TABLE 2 — *The LPG Distribution Parameters* $\alpha = 1, \beta = 1, \theta = .8, \lambda = 1$

x	$h(x)$	$H(x)$
0	.5000	.5000
1	.1243	.6243
2	.0806	.7049
3	.0589	.7638
4	.0451	.8089
5	.0355	.8444
6	.0283	.8727
7	.0228	.8955
8	.0185	.9140
9	.0151	.9291
10	.0124	.9415
11	.0101	.9516
12	.0083	.9599
13	.0069	.9668
14	.0057	.9725
15	.0047	.9772
16	.0039	.9811
17	.0032	.9843
18	.0027	.9869
19	.0022	.9892
20	.0018	.9910

4. A RECURSIVE FORMULA FOR THE DETERMINATION OF THE LPG DENSITY

For numerical calculations, we found the following recursion to be useful. Let us assume that α is integer, and let $C_1 = (\beta/(\lambda + \beta))^\alpha$ and $C_2 = c/(\lambda + \beta)$. Further, using (13) let us define

$$T_{x,k} = C_1 \cdot \frac{\theta^x}{x!} C_2^k S_{x,k} \frac{(\alpha + k - 1)!}{(\alpha - 1)!} \tag{15}$$

Hence, $h(x)$ may be computed as a sum of $T_{x,k}$:

$$h(x) = \sum_{k=1}^{x} T_{x,k}. \tag{16}$$

Note that $T_{x,0} = 0$ since $S_{x,0} = 0$ and that

$$T_{x,x} = C_1 \cdot \frac{(\theta C_2)^x}{x!} \frac{(\alpha + x - 1)!}{(\alpha - 1)!}$$

$$= (\theta C_2/x(\alpha + x - 1)\ T_{x-1,x-1} \tag{17}$$

since $S_{x,x} = 1$. Using (8), we may now write $T_{x,k}$ in terms of $T_{x-1,k}$; specifically, we obtain

$$T_{x,k} = \frac{\theta}{x}[C_2(\alpha + k - 1)T_{x-1,k-1} + (x - 1)\ T_{x-1,k}]. \tag{18}$$

Thus, $h(x)$ may be evaluated using only $T_{x-1,k}$ terms. This provides significant reductions in computer memory and calculation requirements compared to a direct evaluation of (14) for each x.

5. APPROXIMATIONS

Many inventory models require computing reorder points from fractiles of the leadtime demand distribution. Finding exact fractiles of the LPG distribution might be too demanding computationally for many real applications. In this section we consider an approximation which uses a scaled version of the Poisson distribution to approximate the negative binomial distribution.

The mean and variance of $Z(t)$, the number of units demanded in time t, are, respectively,

$$E(Z(t)) = ct\theta/(1 - \theta). \tag{19}$$

$$\text{VAR}(Z(t)) = ct\theta/(1 - \theta)^2, \tag{20}$$

which gives $\text{VAR}(Z'(t))/E(Z(t)) = (1 - \theta)^{-1}$. In certain circumstances, one may have knowledge of the variance to mean ratio of the demand which can then be used to estimate θ directly.

The approximation is based upon replacing the negative binomial distribution of $Z(t)$ with a scaled Poisson distribution. Let Y be a Poisson random variable with parameter μt and let W be defined by $W = kY$ for some $k > 0$. We may think of W as a random variable which assumes values $0, k, 2k, \ldots$ and whose distribution depends upon the two parameters μt and k.

Since

$$E(W) = k\mu t \tag{21}$$

$$\text{VAR}(W) = k^2\mu t \tag{22}$$

we have $\text{VAR}(W)/E(W) = k$. Thus, we set $k = (1 - \theta)^{-1}$ to achieve the same variance to mean ratio. Comparing the mean and variances of W and $Z(t)$ we see that $\mu = c\theta$ (recall that $c = -\lambda/\ln(1 - \theta)$).

Since the negative binomial distribution is defined on all nonnegative integers, we would like the approximation to be defined on the nonnegative integers as well. We have found the following procedure works well. Assume that the scaled Poisson probabilities are shifted to $k/2, 3k/2, 5k/2, \ldots$ so that

$$P\{W = (n+1)\,k/2\} = \frac{e^{-\mu}\mu^n}{n!} \quad n = 0,\ 1,\ 2,\ \ldots. \tag{23}$$

We then assume that the cumulative distribution function is linear between $nk/2$ and $(n+1)k/2$. As an example, suppose $\theta = .75$, $t = 1$, $c = 2$ (that is $\lambda = 2.77$). Then $\mu = 1.5$, $k = 4$ and

$$\begin{aligned} P\{W = 2\} &= e^{-\mu t} = .2231 \\ P\{W = 6\} &= e^{-\mu t}\mu t = .3347 \\ P\{W = 10\} &= e^{-\mu t}(\mu t)^2 = .2510 \\ P\{W = 14\} &= e^{-\mu t}(\mu t)^3/3! = .1255 \\ &\text{etc.} \end{aligned}$$

The comparison of the exact negative binomial probabilities and the scaled Poisson approximation is presented in Table 3 for this case.

TABLE 3 — *Comparison of Negative Binomial and Scaled Poisson Approximations* ($\theta = .75$, $t = 1$, $c = 2$, $\lambda = 2.77$)

x	Negative Bionomial Probabilities	Negative Binomial Cumul. Probabilities	Scaled Poisson Cumul. Probabilities	Scaled Poisson Probabilities
0	.0625	.0625		.0744
1	.0938	.1563		.0744
2	.1055	.2618	.2231	.0744
3	.1055	.3673		.0837
4	.0989	.4662		.0837
5	.0890	.5552		.0837
6	.0779	.6331	.5578	.0837
7	.0667	.6998		.0628
8	.0563	.7561		.0628
9	.0469	.8030		.0628
10	.0387	.8417	.8088	.0628
11	.0312	.8729		.0314
12	.0257	.8986		.0314
13	.0208	.9134		.0314
14	.0167	.9361	.9343	.0314
15	.0134	.9495		.0118
16	.0106	.9601		.0118
17	.0085	.9685		.0118
18	.0667	.9753	.9814	.0118
19	.0053	.9806		.0035
20	.0040	.9846		.0035
21	.0033	.9879		.0035
22	.0026	.9905	.9955	.0035

We now obtain an approximation to the LPG distribution by averaging the scaled Poisson approximation of the negative binomial with the gamma distribution of leadtime. That is,

$$P\{Z(\tau) = x\} \simeq \int_0^\infty \frac{e^{-\mu\tau}(\mu\tau)^{x/k}}{(x/k)!} \cdot \frac{\beta^\alpha \tau^{\alpha-1} e^{-\beta\tau}}{\Gamma(\alpha)}\, d\tau. \tag{24}$$

But this integral is exactly a Poisson mixture with a gamma distribution which is still another way that the negative bionomial distribution can be derived (see Baswell and Patil [2]). Hence, the approximation for the LPG distribution is a scaled version of the negative binomial distribution. The approximation therefore is:

$$P\{Z(\tau) = kx\} \simeq \frac{(\alpha + x - 1)!}{x!\,(\alpha - 1)!}\left(\frac{\beta}{\mu + \beta}\right)^{\alpha}\left(\frac{\mu}{\mu + \beta}\right)^{x} \text{ for } x = 0,\ 1,\ 2,\ \ldots . \tag{25}$$

Note that these probabilities are defined on 0, k, $2k$, ... As with the scaled Poisson we suggest shifting these probabilities to $k/2$, $3k/2$, ... and approximating the probability function by assuming the cumulative distribution function is linear between these fractile points. We tested a variety of cases and found the fit to be excellent, especially in the tails. In Table 4 we compare the exact LPG probabilities for the parameter set considered in Table 2 with the scaled negative binomial approximation. Note that since $\theta = .8$, we have $k = 5$ and the approximate cumulative probabilities (labelled $\bar{H}(x)$ in the table) are defined at the points 2.5, 7.5, 12.5, etc. The final column gives the approximate cumulative distribution function defined on the positive integers obtained from a linear interpolation between the fractiles. Notice the close agreement between the exact and approximate cumulative probabilities in the tail of the distribution.

TABLE 4 — *The Scaled Negative Binomial Approximation to the LPG Distribution*
(Parameters are the same a those of Table 2)

x	Exact Probabilities		Approximate Probabilities		
	$h(x)$	$H(x)$	$\bar{H}(x)$	$h(x)$	Approximate Cumulative
0	.5000	.5000		.1905	.1905
1	.1243	.6243		.1905	.3810
2	.0806	.7049	.6667	.1905	.5715
3	.0589	.7638		.1174	.6889
4	.0451	.8089		.0444	.7333
5	.0355	.8444		.0444	.7777
6	.0283	.8727		.0444	.8221
7	.0228	.8955	.8889	.0444	.8665
8	.0185	.9140		.0296	.8961
9	.0151	.9291		.0148	.9109
10	.0124	.9415		.0148	.9257
11	.0101	.9516		.0148	.9405
12	.0083	.9599	.9630	.0148	.9553
13	.0069	.9668		.0099	.9652
14	.0057	.9725		.0049	.9701
15	.0047	.9775		.0049	.9750
16	.0039	.9811		.0049	.9799
17	.0032	.9843	.9877	.0049	.9848
18	.0027	.9869		.0033	.9881
19	.0022	.9892		.0016	.9897
20	.0018	.9910		.0016	.9913

6. THE FIRST FOUR MOMENTS OF THE LPG DISTRIBUTION

Knowledge of the moments of a complex distribution can be utilized in a variety of ways. The moments can be used to estimate the distribution parameters or to approximate the distribution itself. We derive the first four central moments (moments about the mean) of the LPG distribution.

The distribution of $Z(t)$, the number of units demanded in time t, is negative binomial with parameters $q = \theta$, $p = 1 - \theta$ and $n = ct$. From Kendall and Stuart [7], the first four cumulants of the negative binomial distribution are given by

$$K_1 = nq/p,\ K_2 = nq/p^2,\ K_3 = nq(1 + q)/p^3 \text{ and } K_4 = nq(1 + 4q + q^2)/p^4.$$

The first three cumulants are equal to the first three central moments, respectively, while the fourth central moment, μ_4 is given by

$$\mu_4 = K_4 + 3K_2^2.$$

Hence, the first four central moments (f.f.m.) of $Z(t)$, say μ_i, $1 \leqslant i \leqslant 4$, are

$$\mu_1 = ct\theta/(1 - \theta) \tag{26}$$

$$\mu_2 = ct\theta/(1 - \theta)^2 \tag{27}$$

$$\mu_3 = ct\theta(1 + \theta)/(1 - \theta)^3 \tag{28}$$

$$\mu_4 = ct\theta(1 + 4\theta + \theta^2 + 3ct\theta)/(1 - \theta)^4 \tag{29}$$

In order to derive the f.f.m. of the LPG distribution, we use the following relationships which can be found in Parzen [8], p. 55: Let X and Y be two (dependent) random variables. Then

$$E(Y) = E[E(Y|X)] \tag{30}$$

$$\text{VAR}(Y) = E[\text{VAR}(Y|X)] + \text{VAR}[E(Y|X)] \tag{31}$$

$$\mu_3(Y) = E[\mu_3(Y|X)] + \mu_3[E(Y|X)] \tag{32}$$

$$\mu_4(Y) = E[\mu_4(Y|X)] + 6E[\text{VAR}(Y|X)] \cdot \text{VAR}[E(Y|X)] + \mu_4[E(Y|X)] \tag{33}$$

where

$$\mu_3(Y) = E[(Y - E(Y))^3] \tag{34}$$

$$\mu_4(Y) = E[[Y - E(Y)]^4] \tag{35}$$

In the context of our problem, we interpret Y as $Z(\tau)$ and X as τ. It follows that

$$\begin{aligned} E(Z(\tau)) &= E[E[Z(\tau)|\tau]] \\ &= E[c\tau\theta/(1 - \theta)] \\ &= c\alpha\theta/\beta(1 - \theta). \end{aligned} \tag{36}$$

Similarly,

$$\begin{aligned} \mathrm{VAR}(Z(\tau)) &= E[\mathrm{VAR}(Z(\tau)|\tau] = \mathrm{VAR}[E(Z(\tau)|\tau)] \\ &= E[c\tau\theta/(1-\theta)^2] + \mathrm{VAR}[c\tau\theta/(1-\theta)] \\ &= c\alpha\theta/(\beta(1-\theta)^2) + (c\theta)^2\,\alpha/(\beta^2(1-\theta)^2) \\ &= c\alpha\theta/[\beta(1-\theta)]^2\,\{\beta + c\theta\}. \end{aligned} \tag{37}$$

Following the same kinds of arguments, one eventually obtains

$$\mu_3(Z(\tau)) = \frac{c\alpha\theta}{[\beta(1-\theta)]^3}\,\{\beta^2(1+\theta) + 2c^2\theta^2\} \tag{38}$$

$$\begin{aligned} \mu_4(Z(\tau)) = \frac{c\alpha\theta}{[\beta(1-\theta)]^4}\,&\{\beta^3(1 + 4\theta + \theta^2) + \beta^2 c\theta(3\alpha + 1) \\ &+ 6\beta c^2\theta^2\alpha + c^3\theta^3(3\alpha + 6)\}. \end{aligned} \tag{39}$$

These results can be used to derive exact expressions for the skewness and kurtosis of the LPG distribution. In particular the skewness, say SK, is given by

$$SK = \mu_3(Z(\tau))/[\mathrm{VAR}(Z(\tau))]^{3/2}$$

and the kurtosis (KU) is given by

$$KU = \mu_4(Z(\tau))/[\mathrm{VAR}(Z(\tau))]^2.$$

In Table 5 we present examples of the relationship between the mean, variance, skewness and kurtosis for a variety of system parameters. We can see that each parameter has a different effect on the four measures considered. The mean and the variance both increase with λ, α and θ (although at different rates) and decrease with β. The skewness appears to be most sensitive to changes in α and β while the kurtosis to changes in *alpha,* β and λ and is relatively insensitive to changes in θ.

TABLE 5 — *Means, Variances, Skewness, and Kurtosis for Various System Parameters*

α	β	θ	λ	Mean	Var	SK	KU
1	1	.8	1	2.49	18.60	1.78	20.94
1	1	.8	5	12.43	216.56	1.38	7771.43
1	1	.8	10	24.85	741.96	1.57	84723.23
2	1	.8	1	4.97	37.21	1.26	76.94
5	1	.8	1	12.43	93.02	.79	789.73
10	1	.8	1	24.85	186.04	.56	5393.6
1	2	.8	1	1.24	7.76	2.77	54.77
1	5	.8	1	.50	2.73	5.01	169.02
1	10	.8	1	.25	1.30	7.53	366.37
1	1	.1	1	1.05	2.28	1.09	3.34
1	1	.3	1	1.20	3.16	1.18	9.99
1	1	.5	1	1.44	4.97	1.32	20.35
1	1	.9	1	3.91	54.36	2.15	17.12
1	1	.95	1	6.34	167.07	2.52	16.85
1	1	.99	1	21.50	2611.9	3.35	22.89

ACKNOWLEDGMENT

We are grateful to an anonymous referee for helpful suggestions on an earlier draft.

BIBLIOGRAPHY

[1] Archibald, B.C. and E.A. Silver, "(*s,S*) Policies Under Continuous Review and Discrete Compound Poisson Demand," Management Science , **24**, 899-909 (1978).

[2] Boswell, M.T., and G.P. Patil, "Chance Mechanisms Generating the Negative Binomial Distributions," in *Random Counts in Scientific Work*, Vol. 1, G.P. Patil, Editor, 3-22, (The Pennsylvania State University Press, University Park, 1971).

[3] Chakravarti, I.M., R.G. Laha, and J. Roy, *Handbook of Methods of Applied Statistics*, Vol. 1, (John Wiley and Sons, New York, 1967).

[4] Fisher, R.A., A.S. Corbet, and C.B. Williams, "The Relation Between the Number of Species and the Number of Individuals," Journal of Animal Ecology, *12*, p. 42 (1943).

[5] Hadley, G. and T.M. Whitin, *Analysis of Inventory Systems*, (Prentice Hall, Englewood Cliffs, New Jersey, 1963).

[6] Hausman, W.H., "Minimizing Customer Line Items Backordered in Inventory Control," Management Science, *15*, B-628-B-634 (1969).

[7] Kendall, M.G. and A. Stuart, *The Advanced Theory of Statistics*, Vol. 1, Second Edition, (Hafner, New York, 1963).

[8] Parzen, E., *Modern Probability Theory and Its Applications*, (John Wiley and Sons, New York, 1960).

[9] Presutti, V.J. and R.C. Trepp, "More Ado About EOQ," Naval Research Logistics Quarterly, *17*, 243-251 (1970).

[10] Quenouille, M.H., "A Relation Between the Logarithmic, Poisson and Negative Binomial Series," Biometrics, *5*, 162-164 (1949).

[11] Sherbrooke, C.S., "Discrete Compound Poisson Processes and Tables of the Geometric Poisson Distribution," Naval Research Logistics Quarterly, *15*, 189-204 (1969).

THE OBSERVED HAZARD AND MULTICOMPONENT SYSTEMS*

Mark Brown

Department of Statistics
Florida State University
Tallahassee, Florida

Sheldon M. Ross

Department of Industrial Engineering
and Operations Research
University of California, Berkeley

ABSTRACT

Let X denote the life of some system. We define the observed hazard rate at time t, call it $R(t)$, as the instantaneous probability (density) of failure of X at time t given survival up to t and given a complete description of the system state at t. We conjecture that the total observed hazard—namely, $\int_0^X R(t)\,dt$—is an exponential random variable with mean 1 and verify this for the special case when X is the distribution of system life of an n component system having an arbitrary monotone structure function.

1. INTRODUCTION

Let X denote the survival time of some item and suppose X has distribution function F and density f. Then, the hazard rate function of X—call it $\lambda(t)$—is defined by

$$\lambda(t) = \frac{f(t)}{1 - F(t)}.$$

As

$$\lambda(t)\,dt \approx P\{t < X < t + dt \mid X > t\},$$

we can interpret $\lambda(t)$ as the instantaneous probability (density) that an item of age t will fail.

The distribution F can be expressed as

$$1 - F(t) = \exp\{-\int_0^t \lambda(s)\,ds\},$$

implying that

$$1 - F(X) = \exp\{-\int_0^X \lambda(s)\,ds\}.$$

*This research has been partially supported by the Office of Naval Research under Contract N00014-77-C-0299 and the Air Force Office of Scientific Research (AFSC), USAF, under Grant AFOSR-77-3213B with the University of California.

Now, as is well-known, $1 - F(X)$ has a uniform distribution on (0,1) and as the negative logarithm of such a random variable has an exponential distribution with mean 1, it follows that

$$\int_0^X \lambda(s)\,ds \sim \text{Exponential (1)},$$

or, in words, the total hazard experienced by the item is exponentially distributed with mean 1.

In the above, the hazard rate at time t was defined to be the probability (density) of failure at t given survival up to that time. Now, however, let us suppose that we define the observed hazard at time t—call it $R(t)$—to be a random variable which represents the actual probability (density) of death at time t given not only the fact of survival up to time t but also a complete description of the "state" of the item at that time. (Such a quantity would, in general, be a random variable as it would be a function of the "state" of the item at time t and the state would itself generally be a random variable.) For instance, for a given individual, $R(t)$ would denote the probability (density) of failure at time t given the life history of the individual up to time t.

We conjecture that $\int_0^X R(t)\,dt$ is also an exponentially distributed random variable and in the following section, we verify this conjecture in the case of an n component coherent system in which components function for a random time and then fail.

2. COHERENT SYSTEMS AND THE OBSERVED HAZARD

We are given a system consisting of n components each of which is at all times either working or failed. In addition, we suppose that whether or not the system is working is solely determined as a function function—call it ϕ—of the component states. That is, letting x_i equal 1 or 0 according to whether or not the i^{th} component is working, then we suppose that there exists a nondecreasing binary function ϕ such that

$$\phi(\underline{x}) = \phi(x_1, \ldots, x_n) = \begin{cases} 1 \text{ if system works under state vector } \underline{x} \\ 0 \text{ otherwise} \end{cases}$$

Suppose now that component i is initially working and will work for a random time having distribution F_i at which time it will fail, $i = 1, \ldots, n$. Once a component has failed, it remains failed from that time on. Let $x_i(t)$ equal 1 if component i is working at time t and 0 otherwise and define the random sets $C(t)$ by

$$C(t) = \{i : \phi(1_i, \underline{x}(t)) = 1, \phi(0_i, \underline{x}(t)) = 0\}$$

where $\phi(1_i, \underline{x}) = \phi(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n)$ and $\phi(0_i, \underline{x}) = \phi(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n)$. In words, $C(t)$ is the set of critical components at time t, where a component is critical at some time if its failure at that time would cause the system to go from working to failed.

If we let L denote the length of system life, then assuming independence of components, the total observed hazard experienced by the system during its lifetime could be expressed as

$$\int_0^L \sum_{i \in C(t)} \lambda_i(t)\,dt = \text{total system hazard}$$

where $\lambda_i(t)$ is the (usual) hazard rate function of component i at time t. We now show that this random quantity is exponentially distributed with mean 1.

THEOREM: The total system hazard is an exponential random variable having mean 1.

PROOF: The proof is by induction on n. As the observed hazard rate is equal to the (usual) component hazard rate when $n = 1$, the result follows in this case. So assume the result for any system of $n - 1$ components and consider an n component system. Say a component is a 1-component minimal cut set if its failure guarantees system failure even if all other components are working. We consider three cases.

CASE 1: There do not exist any 1-component minimal cut sets.

In this case, the observed hazard rate will be 0 until a component fails. At this point, the remaining hazard will be exponentially distributed with mean 1 by the induction hypothesis. Hence, the result follows in this case.

CASE 2: There exists exactly one 1-component minimal cut set—say {1}.

In this case, let L_1 denote the life of component 1 and let T denote the first time any of the components 2 through n fail. Now, conditional on $T = t$, the total observed hazard can be expressed as

$$\int_0^{L_1} \lambda_1(s)\,dt \text{ if } \int_0^{L_1} \lambda_1(s)\,ds \leqslant \int_0^{t} \lambda_1(s)\,ds$$

$$\int_0^{t} \lambda_1(s)\,ds + \text{Exp (1) if } \int_0^{L_1} \lambda_1(s)\,ds > \int_0^{t} \lambda_1(s)\,ds$$

where we have used the induction hypothesis in writing that the remaining observed hazard starting at time $T = t$ and assuming that component 1 has not yet failed is exponentially distributed with mean 1. Thus, from the above, we see that, given $T = t$, the total observed hazard has the same distribution as the random variable defined by

$$E_1 \text{ if } E_1 < c$$

$$c + E_2 \text{ if } E_2 > c$$

where c is a constant and E_1, E_2 are independent exponential random variables each having mean 1. Such a random variable is easily seen to be also exponential with mean 1.

CASE 3: There exists at least two 1-component minimal cut sets—say {1} and {2}.

In this case, we can combine components 1 and 2 into a single component which fails when either one of them fails and the result follows from the induction hypothesis.

REMARK: The above proof goes through in an identical manner even when the component lifetimes are dependent. Of course, the observed hazard rate at time t would no longer be $\sum_{i \in C(t)} \lambda_i(t)$ but would have to be suitably modified.

3. SOME FINAL REMARKS AND A HEURISTIC ARGUMENT

(i) Whereas we have only established that the total observed hazard experienced by a system is exponentially distributed with mean 1 for the rather special system described in Section 1, we believe that this result holds with tremendous generality. (Another system in which we have been able to verify it is when events

occur in accordance with some arbitrary point process and each event has a random nonnegative damage associated with it. The system is said to fail the first time the total cumulative damage exceeds some specified value.)

(ii) An interesting sidelight about the system of Section 1 is that it is well-known that if all component life distributions are IFR (increasing failure rate) in the sense that $\lambda_i(t)$ is a monotone nondecreasing function for all $i = 1, \ldots, n$, then it need not be the case that the system hazard rate is also increasing. However, it easily follows in this case of IFR component life distributions that the observed system hazard rate—namely, $\sum_{i \in C(t)} \lambda_i(t)$—increases up to the time of system failure.

(iii) A general definition of the random hazard can be given along the following lines: Let $\{F_s, 0 \leqslant s <= \infty\}$ denote an increasing family of sigma fields and let X denote a stopping time defined on this family. Let

$$R(t) = \lim_{h \to 0} \frac{P\{t < X < t + h \mid F_t\}}{h}$$

where we assume the above limit exists almost surely. We now claim that, when it is well defined,

(2.1) $\int_0^X R(t)\,dt$ has an exponential distribution with mean 1.

We now present a heuristic argument of the above.

Heuristic Argument of (2.1)

Let $H(t) = \int_0^t R(s)\,ds$. We wish to argue that $H(X)$ is exponential with mean 1 and to do so, we shall argue that its failure rate function—call it $\lambda(s)$—is identically 1. To show that $\lambda(s) \equiv 1$, let us condition on the event that $H(X) \geqslant s$ and on the values of T_s and $R(T_s)$ where T_s is defined to be the time at which H is equal to s—that is, $H(T_s) = s$. Now given $H(X) \geqslant s$, T_s, $R(T_s) = \lambda_s$

$$X \geqslant T_s + \epsilon \text{ with probability } 1 - \epsilon\lambda_s + o(\epsilon).$$

Hence,

$$H(X) \geqslant H(T_s + \epsilon) \text{with probability } 1 - \epsilon\lambda_s + o(\epsilon).$$

But

$$\begin{aligned} H(T_s + \epsilon) &= H(T_s) + \epsilon\lambda_s + o(\epsilon) \\ &= s + \epsilon\lambda_s + o(\epsilon) \end{aligned}$$

and so

$$H(X) \geqslant s + \epsilon\lambda_s + o(\epsilon) \text{with probability } 1 - \epsilon\lambda_s + o(\epsilon)$$

which, for $\lambda_s > 0$, is roughly equivalent to

$$H(X) \geqslant s + \delta \text{ with probability } 1 - \delta + o(\delta).$$

Thus, independent of T_s, λ_s, given that $H(X) \geqslant s$, it has probability $\delta + o(\delta)$ of failing during the next δ units of hazard. Thus,

$$P\{H(X) < s + \delta | H(X) \geqslant s\} = \delta + o(\delta).$$

Dividing the above by δ and letting δ go to 0 "proves" that the failure rate function of $H(X)$ is identically one.

A REPLACEMENT SCHEDULE FOR MULTICOMPONENT LIFE-LIMITED PARTS

Sheldon Epstein

Department of Quantitative Analysis
Seton Hall University
South Orange, New Jersey

Yonah Wilamowsky

Department of Quantitative Analysis
Long Island University
Brooklyn, New York

ABSTRACT

In previous papers [1], [2] the authors developed a maintenance policy for a single life-limited part. Using an *opportunistic replacement* approach a scheme was devised which utilized early replacement of equipment to offset more costly future expenditures. This paper will extend the results to the multicomponent case. Examples are given illustrating the benefits of this new technique.

INTRODUCTION

Optimal opportunistic replacement (OR) and maintenance policies for single component systems have been extensively discussed [3], [4], [5]. In previous papers [1], [2], the authors dealt with the application of these policies to life-limited parts. Specifically, B is a component of a system with the following properties:

(1) The system is known to fail exponentially (not due to B) at which time it is brought to the maintenance base at a cost of C_T; disassembled and reassembled (if necessary) at a cost of C_A (cost of accessing); and repaired according to its malfunction.

(2) B never fails but has a life-limit of Y hours. After Y hours, the machine must be brought to the maintenance base (forced removal—FR) at a cost of C_T, disassembled and reassembled at a cost of C_A ($C_F = C_T + C_A$) and part B replaced at a cost of C_B.

Compressor units on jet engines, for example, contain disks that satisfy these conditions. The optimal strategy for B is given by an (X_1, X_2) rule such that:

(1) If, when the unit fails, B is accessible, replace B if it has less than X_2 hours of remaining life.

(2) If, when the unit fails, B is not accessible, replace B if it has less than X_1 hours of remaining life ($X_1 \leqslant X_2$).

(3) Under all conditions replace B after Y hours of usage.

X_1 and X_2 are computed by minimizing

$$\frac{\text{Expected (Cost)}}{\text{Expected (Life)}} = \frac{E(C)}{E(L)}$$

where

(1) $$E(C) = C_B + \exp\left[\frac{-(X_2 - X_1)p}{\lambda}\right]((1-p)(1-e^{-X_1/\lambda})C_A + e^{-X_1/\lambda}C_F)$$

(2) $$E(L) = Y - X_2 + \lambda\left[\frac{1}{p} + \exp\left[-\frac{(X_2 - X_1)p}{\lambda}\right]\left[1 - \frac{1}{p} - e^{-X_1/\lambda}\right]\right]$$

and

λ — Average time between removals (due to the unit failure)

p — Probability of B being accessible on any unit breakdown.

This paper will extend our previous results to a multicomponent system (i.e., more than one life-limited disk). While the above-mentioned policies insure optimal utilization of individual components the results may be far from optimal for the system as a whole.

EXAMPLE 1: Assume two identical disks, D_1 and D_2, with the following parameters, are inserted simultaneously into a jet engine compressor unit (a unit may contain as many as 10 disks):

$$p = 1,\ C_B = \$18{,}000,\ \lambda = 5{,}000 \text{ hours},\ C_F = \$7{,}800, \text{ and } Y = 30{,}000 \text{ hours}.$$

Applying our rule on an individual basis yields $X_1 = 0$ and $X_2 = 5{,}650$. The cost of the entire system per hour is thus, from (1) and (2)

$$\frac{E(C_{D_1})}{E(L_{D_1})} + \frac{E(C_{D_2})}{E(L_{D_2})} = \frac{E(C_{D_1})}{E(L_{D_1})} + \frac{18{,}000}{E(L_{D_2})} = \frac{20{,}520}{27{,}735} + \frac{18{,}000}{27{,}735} = \$1.389 \text{ per hour}.$$

$E(C_{D_2}) = 18{,}000$ because (a) $p = 1$ cancels the C_A term; and (b) no forced removal cost C_F is assigned to this disk since if an FR does occur both disks will be removed simultaneously, causing only one C_F which has already been accounted for in the calculation of $E(C_{D_1})$. The optimal policy for the system is in fact $X_1 = 0$ and $X_2 = 3{,}800$ and costs

$$\frac{21{,}648}{28{,}863} + \frac{18{,}000}{28{,}863} = \$1.373 \text{ per hour}.$$

Considering that major airlines operate hundreds of engines with thousands of disks, a 0.8 cents savings per disk per hour is appreciable.

The solution of the multicomponent system has proven to be elusive. Various enumerative procedures for the general class of OR problems have been suggested, [6], [7]. Application of these techniques to real-life situations, however, requires simplifying assumptions that are

unrealistic and often result in considerably suboptimal policies. Additionally, computational requirements for these techniques are quite large. Vergin and Scriabin [7] state that, on an IBM 370/165, the two disk problem required 0.5 minutes and the three disk problem 13.3 minutes. Any attempted extension to an n-disk problem ($n > 3$) leads to prohibitively high computer processing time and memory requirements. They conclude by saying that "it seems highly unlikely that any computationally simple optimizing model will be developed for the maintenance of very complex equipment."

The structure of the life-limited problem and the technique of our solution in the single disk case offer an approach not previously investigated. It involves an analytical procedure that facilitates handling of even the n-disk paradigm. We present the solution to the two disk problem. The generalization to n-disks is similar. Illustrative examples are provided.

PROBLEM FORMULATION

Consider two disks, 1 and 2, where 2 has ΔT more hours of life remaining than 1 (Figure 1). We will, for the sake of a more lucid development, assume throughout this paper that $p = 1$, i.e., every engine maintenance insures accessibility of the entire unit. The generalization for $p < 1$ can be readily derived. The following notation will be used:

X_i — The removal point for disk i (only one accessible cut-off point per disk is necessary).

C_{B_i} — Cost of disk i.

Y_i — Life limit of disk i.

$E(C_i)$ — Expected cost of disk i in a one disk system.

$E(L_i)$ — Expected life of disk i in a one disk system.

$E(C_{2/1})$ — Expected cost of disk 2 in a system which also contains disk 1.

$E(L_{2/1})$ — Expected life of disk 2 in a system which also contains disk 1.

P_1 — The probability that disk 2 is retained when disk 1 is removed.

$E(C_i)$ and $E(L_i)$ can be found by substituting the desired X_i into the simplified forms of equations (1) and (2), i.e.,

$$E(C_i) = C_{B_i} + e^{-X_i/\lambda} C_F \tag{3}$$

$$E(L_i) = Y_i - X_i + \lambda(1 - e^{-X_i/\lambda}). \tag{4}$$

Any time disk 1 arrives with less than X_1 hours of remaining life it will be removed and disk 2 will be checked with respect to X_2. If 1, however, exceeds X_1, disk 2 under all circumstances will remain. The reasoning is simple. An FR attributable to 2 is impossible if 1 is still operational. In the worst case, 1 will lead to an FR and make 2 accessible at that time. Since removal of a disk is justifiable only as a deterrent to a possible FR, a decision to retain 1 must automatically retain 2 as well. A flowchart of our policy is given in Figure 2.

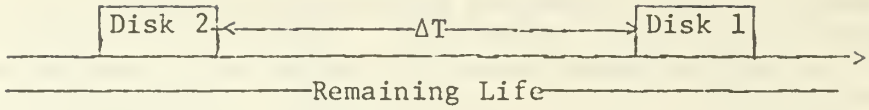


FIGURE 1. Two disk paradigm

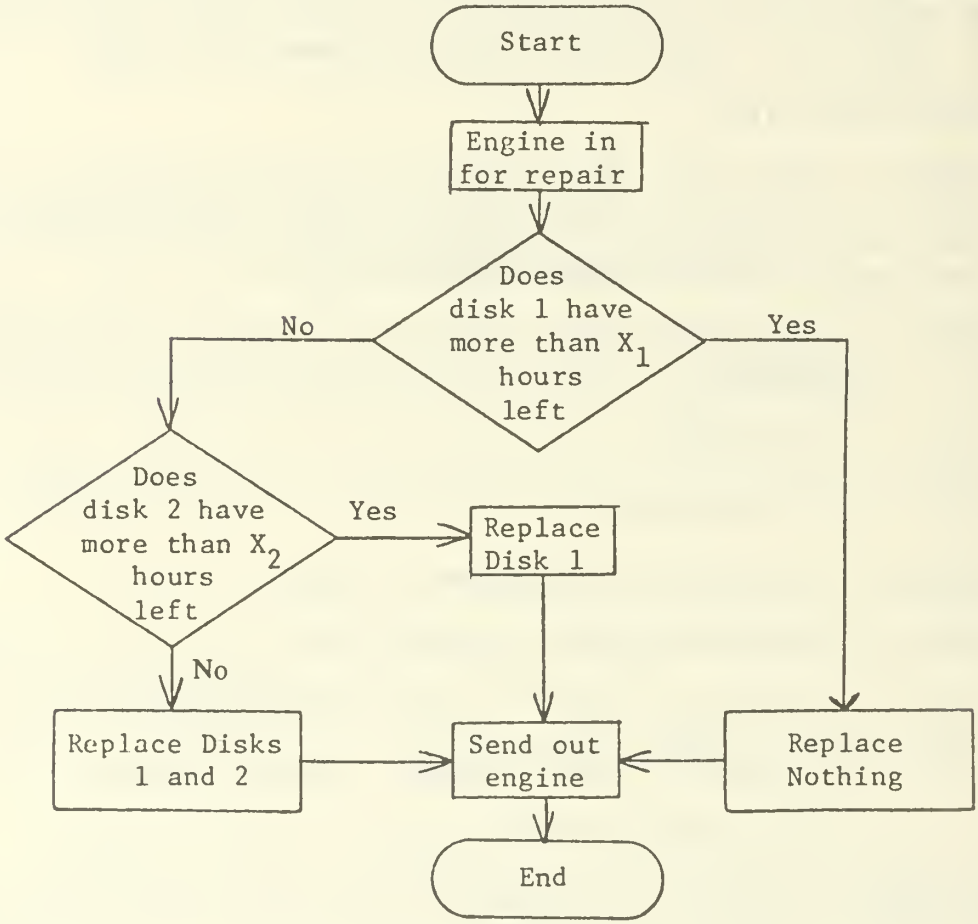


FIGURE 2. Two disk flowchart

ANALYSIS

The relevant factors are summarized in the following series of statements:

STATEMENT 1: The cost per hour of the system is equal to the sum of the costs per hour of the individual disks.

STATEMENT 2: The objective in the 2-disk problem is to minimize

$$\frac{E(C_1)}{E(L_1)} + \frac{E(C_{2/1})}{E(L_{2/1})} \tag{5}$$

PROOF: The cost and life of disk 1 is a function of X_1 alone. The cost and life of disk 2, however, as we have previously explained is dependent on both X_1 and X_2.

STATEMENT 3: The optimal X_2 may be derived by the single disk approach, i.e., find the X_2 that minimizes $E(C_2)/E(L_2)$ of (3) and (4).

PROOF: Since X_2 is used only if disk 1 has been removed, the statement is true by dynamic programming.

STATEMENT 4:

$$E(C_{2/1}) = P_1E(C_2) + (1 - P_1)C_{B_2} \tag{6}$$

PROOF: If disk 2 remains after disk 1 has been removed, then it may generate an FR. Its expected cost is therefore given by $E(C_2)$ of (3). Disk 2 can never cause a forced removal if it is removed together with disk 1. Its expected cost in this situation is simply its original cost.

STATEMENT 5: The expected life of disk 2 given that it is removed together with disk 1 is:

$$Y_2 - \text{Min}\,(X_2,\ X_1 + \Delta T) + \lambda(1 - e^{-\text{Min}(X_2-\Delta T, X_1)/\lambda}). \tag{7}$$

PROOF: (a) If $X_2 \geqslant X_1 + \Delta T$ then whenever disk 1 is below its cut-off value disk 2 is also below its cut-off point. Thus, the two disks will always be removed together. The double removal will occur due to an accessible removal when disk 1 has less than X_1 hours remaining (and disk 2 has less than $X_1 + \Delta T$), or an FR when all of disk 1's permissible life has been expended (and disk 2 has ΔT of remaining life). The expected value is

$$\int_0^{X_1} (Y_2 - (X_1 + \Delta T) + x)\,\frac{1}{\lambda}\,e^{-x/\lambda}dx + (Y_2 - \Delta T)\int_{X_1}^{\infty}\frac{1}{\lambda}\,e^{-x/\lambda}dx.$$

(b) If $X_2 < X_1 + \Delta T$ it is possible that disk 1 is removed and disk 2 is retained. The knowledge that both are removed simultaneously insures that both were in use when disk 2 had X_2 hours of remaining life, or equivalently when disk 1 had $X_2 - \Delta T$ hours of life remaining. Since $X_2 - \Delta T < X_1$ the next breakdown, if it occurs prior to an FR, will result in the removal of both disks. The expected life is therefore,

$$\int_0^{X_2-\Delta T} (Y_2 - X_2 + x)\,\frac{1}{\lambda}\,e^{-x/\lambda}dx + (Y_2 - \Delta T)\int_{X_2-\Delta T}^{\infty}\frac{1}{\lambda}\,e^{-x/\lambda}dx.$$

Simplifying both expressions yields (7).

STATEMENT 6:

$$E(L_{2/1}) = P_1E(L_2) + (1 - P_1)(Y_2 - \text{Min}(X_2,\ X_1 + \Delta T) + \lambda(1 - e^{-\text{Min}(X_2-\Delta T, X_1)/\lambda})). \tag{8}$$

PROOF: Direct application of expected value and (7).

STATEMENT 7:

$$P_1 = \begin{cases} 0 & \text{if } X_1 + \Delta T < X_2 \\ 1 & \text{if } X_2 < \Delta T \\ 1 - e^{-(X_1+\Delta T-X_2)/\lambda} & \text{otherwise} \end{cases} \tag{9}$$

PROOF: In the first case the disks will always be removed together while in the second case the time differential is so great that the disks have no effect on each other. In all other cases a breakdown with ΔX_1 remaining time on disk 1 ($\Delta X_1 = X_1 - x$, where x is time to first breakdown after disk has reached its critical value) means that disk 2 has $\Delta X_1 + \Delta T$ time left.

$$\begin{aligned} P_1 &= P(\Delta X_1 + \Delta T > X_2) \\ &= P(X_1 - x + \Delta T > X_2) \\ &= P(x < X_1 + \Delta T - X_2) \\ &= \int_0^{X_1+\Delta T-X_2} \frac{1}{\lambda} e^{-x/\lambda} dx \end{aligned}$$

which simplifies to (9).

It is now clear that the objective function (5) can be broken down into a function of X_1 and X_2 using equations (6), (8), and (9). Since X_2 has already been computed in the standard manner, the only remaining variable is X_1. The optimal X_1 can thus be computed by using a one dimensional search on the objective function.

EXAMPLE 2: Referring to Example 1 ($\Delta T = 0$), an X_2 value of 5,650 minimizes $E(C_2)/E(L_2)$. Inserting this value into (8) and (9) the X_1 value that minimizes the objective function (5) is 3,800. Note that P_1 for this combination is $0(X_1 + \Delta T < X_2)$, guaranteeing that both disks will always be removed simultaneously and that X_2 need never be actually applied. It is intuitively obvious and can be easily shown that one disk having the combined cost of disks 1 and 2 will have the same 3,800 cut-off value.

EXAMPLE 3: We will again use the 2 disks of Example 1 except that this time disk 1 is 3,000 hours older than disk 2, i.e., $\Delta T = 3{,}000$. As before X_2 is initially computed as 5,650 and an X_1 of 4,250 will minimize (5) at a cost of $\frac{21,334}{28,613} + \frac{18,677}{26,754} = 1.444$ dollars per hour. Because the 2 disks have different remaining lives this system costs an additional 7 cents per hour to run (1.444-1.373). If the engine is in for repairs when

(a) disk 1 has more than 4,250 hours left—leave both disks

(b) disk 1 has less than 4,250 but more than 2,650 hours remaining—remove disk 1 and leave disk 2

(c) disk 1 has less than 2,650 hours left—replace both disks.

EXAMPLE 4: Suppose the two disks differ in age by over 5,650 hours, i.e., $\Delta T > 5{,}650 = X_2$. By (9) $P_1 = 1$ and everything reduces to a computation of the individual optimal cut-off points for each disk, i.e., $X_1 = X_2 = 5{,}650$. The system costs $0.74 + 0.74 = \$1.48$ per hour. We can conclude that as the disk separation time increases from 0 to X_2 the minimal cost per hour will continuously increase.

CONCLUSIONS

A variation of this procedure was used in a simulation run of actual engine configurations of a major passenger airline. Computational requirements caused no problem for even a seven-disk system. Additionally, the results were compared to various alternate approaches (e.g. Markov analysis) and found to be superior to them all.

It should be noted that certain underlying assumptions are inherent in our analysis. In a 2-disk system for example, computation of X_2 is necessary for the computation of X_1 and it is assumed if disk 1 is removed, X_2 will be our criteria for judging the wisdom of removing disk 2. This, however, depends on the amount of life in disk 1's replacement. If the new disk has less life remaining than disk 2, then by previous arguments disk 2 will be retained regardless of X_2. If the new disk, on the other hand, has more remaining life than disk 2, then disk 2 becomes disk 1 in the new configuration. As such, its cut-off point is not the previously computed X_2 but a new X_1. In order for our procedure to yield optimal results the new X_1 would have to equal the old X_2. As discussed in the 2-disk system if $\Delta T > X_2$, disk 1 is unaffected by disk 2 and can be treated as an individual disk. Life specifications of jet engines are such that if disk 2 in the initial system was close enough in life to disk 1 to influence it, then the new disk which replaces disk 1 will have considerably more life than the remaining disk, and hence not affect it. Under these conditions the use of X_2 in our policy is justified.

However, what if we are dealing with a situation where this condition is not met? Engine build policies may very well dictate the replacement of a worn out disk with remaining life that is comparable to the retained disk. As discussed in Example 4 such a policy offers configurations leading to lower costs per hour. However, the necessary prerequisite for guaranteed optimality is missing. Additionally, when removing one or two disks the time differential between the disk will generally change. Hence, although we minimize the present configuration we in no way consider the effect our decision will have on future utilization.

Because of these arguments the multidisk extension is presented not as the optimal strategy but rather as a heuristic. The complexity of the problem necessitates the use of a more limited objective. Disregarding future contingencies, we restrict ourselves to the optimization of currently operating disks only. The application of a logical policy in the present will hopefully lead to near optimal system utilization in the future as well. The simulation runs previously mentioned indicated that our policy will converge on uniform configurations in a relatively short period of time. Pending further investigation, the authors are convinced that, at least in the jet-engine problem, the approach is sound and far superior to any other presently employed.

REFERENCES

[1] Epstein, S. and Y. Wilamowsky, "A Statistical Model for Jet Engine Maintenance," presented at the Annual Meeting of the American Statistical Association, San Diego, CA (August 1978).

[2] Epstein, S. and Y. Wilamowsky, "A Disk Replacement Policy for Jet Engines," The Annals of the Society of Logistics Engineers, *5*, 1 (Summer, 1980).
[3] Jorgenson, D.W. and R. Radner, "Optimal Replacement and Inspection of Stochastically Failing Equipment," RAND, P-2074 (1960).
[4] McCall, J.J., "Solution of a Simple Overhaul Problem," RAND, RM 2989 PR (February, 1962).
[5] Radner, R. and D.W. Jorgenson, "Optimal Replacement of a Single Part in the Presence of Several Monitored Parts," Management Science, *10*, 1 (1963).
[6] Sethi, D.P.S., "Opportunistic Replacement Policies," in *The Theory and Applications of Reliability: With Emphasis on Bayesian and Nonparametric Methods,* Vol. 1, 433-447, C.P. Tsokos and I.N. Shimi, Editors (Academic Press, New York, 1977).
[7] Vergin, R.C. and M. Scriabin, "Maintenance Scheduling for Multicomponent Equipment," AIIE Transactions, *9*, 3 (September, 1977).

A NOTE ON RANDOMLY EVOLVING HAZARD RATE FUNCTIONS*

Zvi Schechner

Department of Industrial Engineering and Operations Research
Columbia University
New York, NY

ABSTRACT

Let τ be a finite stopping time with random hazard rate function $\{\lambda_t : t \geq 0\}$. We prove that $\int_0^\tau \lambda_t \, dt$ is exponentially distributed with mean 1.

1. INTRODUCTION

In their paper published in this issue (Naval Research Logistics Quarterly, *29*, 4, 1982), Brown-Ross [1] introduce the notion of a "randomly evolving hazard rate function." Perhaps the simplest case of such a function is known in reliability theory as the failure rate function. Let X denote the survival time of some item and suppose X has distribution function F and density f. Then, the failure rate function of X—call it $\lambda(\cdot)$—is defined by:

$$\lambda(t) = \frac{f(t)}{1 - F(t)} \tag{1.1}$$

and since $\lambda(t)\, dt \cong P(X \in (t, t + dt] \mid X > t)$ one can interpret $\lambda(t)$ as the instantaneous probability density that an item of age t will fail. It is well known that the total hazard experienced by the item, i.e.,

$$\int_0^X \lambda(t)\, dt \tag{1.2}$$

is exponentially distributed with unit rate. Now suppose that we are able to observe more than just whether or not the item is functioning at t but also a complete description of the "state" of the item at that time. Then the interpretation of $\lambda(\cdot)$ becomes:

$$\lambda(t)\, dt \cong P(X \in (t, t + dt] \mid \text{complete information of state of item at } t). \tag{1.3}$$

Such a quantity is, in general, a random variable and its value is determined by the state of the component at t which is itself a random variable. Brown-Ross [1] conjectured that even in this case $\int_0^X \lambda(t)\, dt$ is exponentially distributed with unit rate. They verified the conjecture in numerous cases taken from well-known models in statistics, reliability and Markov chains. However, for the general case they only provided a heuristic proof.

*This research was supported by the National Science Foundation under Grant MCS-81-02075.

Using recent developments in the martingale theory of jump processes, we shall give a precise formulation of the conjecture and prove it.

2. THE CONJECTURE AND ITS PROOF

Let (Ω, F, P) be a probability space and let $\{H_t : t \geq 0\}$ be a complete, right continuous, increasing family of sub-σ-*fields*. A stochastic process $\{X_t : t \geq 0\}$ is said to be adapted to $\{H_t\}$ if for each $t \geq 0$ X_t is H_t-measurable. It is said to be $\{H_t\}$ predictable if there exists a sequence $\{X_t^{(n)}\ t \geq 0\}$ of stochastic processes such that for each $t \geq 0$ $\lim_n X_t^{(n)} = X_t$ and for each n $\{X_t^{(n)}\}$ is left continuous and adapted to $\{H_t\}$.

Let τ be any finite stopping time defined on $\{H_t\}$, i.e., $P(0 \leq \tau < \infty) = 1$ and for each $t \geq 0$, $\{\tau \leq t\}$ belongs to H_t. Define the following process:

$$I_t = \begin{matrix} 1 & \text{if } \tau \leq t \\ 0 & \text{otherwise} \end{matrix} \tag{2.1}$$

clearly I_t is a submartingale. Hence, by the Doob-Meyer decomposition theorem there exists a unique increasing process, call it $\{A_t : t \geq 0\}$, such that:

(2.2) (i) $A_0 = 0$

(ii) A is right continuous

(iii) A is $\{H_t\}$-predictable

and:

(2.3) $M_t = I_t - A_t$ is $\{H_t\}$ martingale. Now suppose that $\{A_t\}$ has a differentiable sample path, and let $\lambda_t = \frac{d}{dt} A_t$. Then the stochastic process $\{\lambda_t : t \geq 0\}$ is the random hazard rate of the stopping time τ. To see this, note that since $\{M_t\}$ is a martingale, we have:

$$E[M_{t+dt} - M_t | H_t] = 0$$

hence

$$E[I_{t+dt} - I_t | H_t] = E[A_{t+dt} - A_t | H_t]$$

but

$$E[I_{t+dt} - I_t | H_t] = P(\tau \in (t, t + dt] | H_t)$$

and

$$E[A_{t+dt} - A_t | H_t] \cong \lambda_t\, dt.$$

Now we state the conjecture as a theorem:

THEOREM: If $\{A_t\}$ is differentiable with derivative $\{\lambda_t\}$ then $A_\tau \equiv \int_0^\tau \lambda_t\, dt$ has an exponential distribution with mean 1.

The proof is based on the following lemma:

LEMMA: $E[A_\tau^n] = n!\quad n = 0,1,2,\ \cdots$.

PROOF: It is well known (consult for instance [3. Ch. 18]) that for any finite $\{H_t\}$-predictable process $\{ f_t : t \geqslant 0\}$ the stochastic integral $\int_0^t f_u\, d\, M_u$ is a $\{H_t\}$-martingale. In particular, since $\{A_t\}$ is predictable, for each n $(n = 0,1,2..)$

$$q_t^{(n)} = \int_0^t A_u^n\, dM_u \equiv \int_0^t A_u^n\, dI_u - \int_0^t A_u^n \lambda_u\, du \tag{2.4}$$

is a martingale.

Thus: $E[q_\tau^{(n)}] = 0$

$$E\left[\int_0^\tau A_u^n\, dI_u\right] = E\left[\int_0^\tau A_u^n \lambda_u\, du\right] \tag{2.5}$$

or equivalently:

$$[A_\tau^n] = E\left[\frac{A_\tau^{n+1}}{n+1}\right] \tag{2.6}$$

which implies that

$$[A_\tau^n] = n!\quad n = 0,1,\ \cdots . \tag{2.7}$$

This completes the proof of the Lemma.

PROOF OF THE THEOREM: First note that the nth moment of an exponential distribution with mean 1 is $n!$

It is well known [2. p.228] that a distribution function is uniquely determined by its moments $\{\mu_k\}$ whenever the power series:

$$\sum_n \frac{\mu_{2n}\, t^n}{(2n)!}$$

converges in some interval. This clearly holds for the exponential distribution and thus A_τ has an exponential distribution with mean 1. This completes the proof of the theorem.

REMARK: Note that a sufficient condition for A_τ to have an exponential distribution with mean 1 is that the sample path of A is continuous, (2.4) becomes $\int_0^t A_u^n\, dI_u - \int_0^t A_u^n\, dA_u$ which for each sample path of A is merely a difference between two Stieltjes integrals and thus (2.6) and (2.7) remain valid.

BIBLIOGRAPHY

[1] Brown, M. and S.M. Ross, "The Observed Hazard and Multicomponent Systems," Naval Research Logistics Quarterly, *29*, 4 (1982).

[2] Feller, W., *An Introduction to Probability Theory and Its Applications,* Vol. II, (John Wiley and Sons, New York, 1971).

[3] Lipster, R.S. and A.N. Shiryayev, *Statistics of Random Processes II: Applications,* (Springer Verlag, New York, 1978).

[4] Varaiya, P., "The Martingale Theory of Jump Processes." IEEE Transactions on Automatic Control, *AC-20*, 1, 34-42 (1975).

INDEX TO VOLUME 29

FUTURE PUBLICATION OF NAVAL RESEARCH LOGISTICS QUARTERLY

The Office of Naval Research will discontinue publication of the Naval Research Logistics Quarterly after distribution of the December 1982 issue.

The journal will continue to be published, however, and will appear as a nongovernmental periodical to be published with the cooperation of the Office of Naval Research.

The new publisher will be John Wiley and Sons, Inc. The Naval Research Logistics Quarterly will become a copyrighted journal in the Wiley-Interscience series, and will be sold on a subscription basis by the publisher. The cost for a one year subscription will be $60.00 to individuals and institutions. Wiley-Interscience publication will be initiated with the March 1983 issue, Vol. 30, No. 1.

A substantial professional discount will be offered to members of the Operations Research Society of America and The Institute for Management Sciences. Members of ORSA or TIMS may subscribe to the journal for $20.00 per year.

If you are currently receiving the NRLQ by paid subscription through the Superintendent of Documents, U.S. Government Printing Office, and your subscription is due to expire after the December 1982 issue, a refund for the unfulfilled portion of your subscription will be issued to you by the Government Printing Office. The new publisher has agreed to honor requests at the GPO rate for unfulfilled portions which were subscribed to previously. The Superintendent of Documents will notify subscribers if they are due refunds because of such termination, and refund checks will be mailed subsequently. Subscribers may then elect to extend their subscriptions at the GPO rate by contacting the publisher and enclosing their endorsed refund checks in payment for the issues due subsequent to termination. An order form to facilitate this extension and/or renewal by the new publisher is enclosed.

Distribution through officially authorized Navy lists and SupDocs dissemination to government repository library centers will be discontinued after the December 1982 issue.

The Office of Naval Research regrets that it cannot provide support incident to such distribution and that those activities requiring copies for official purposes will be expected to bear the necessary costs directly.

Due to restrictions governing the use of GPO subscription lists, the publisher will be unable to follow-up with solicitations mailed with listings provided by the Superintendent of Documents; hence, future promotional efforts may only be through professional and trade sources. Because of this some current recipients may not receive anticipated additional notices from the publisher, and may not maintain the continuity of their collections if they rely upon the receipt of subscription solicitations from the publisher. Thus, all present recipients who are interested in receiving the journal are encouraged to initiate communications with the publisher promptly so that their names will not be purged by default.

It is expected that current editorial policies will be continued in the commercially published version. Professor Herbert Solomon of Stanford University will be the new editor in charge and will be responsible for the review of manuscripts in the future. New manuscripts submitted for consideration for publication may be sent to him immediately (4 copies) at the following address:

Professor Herbert Solomon, Editor
Naval Research Logistics Quarterly
Department of Statistics
Sequoia Hall
Stanford University
Stanford, California 94305

The Office of Naval Research is pleased with the recognition of the contributions to research promulgated through the Naval Research Logistics Quarterly. As ONR's own scientific journal it has reflected Navy interests in diverse areas of mathematics research which we believed would enrich future work in logistics and systems analysis, and has well served its intended purpose of providing a forum for mathematicians and logistics engineers to interact at the highest scientific level. The willingness of the publisher and scientific community to perpetuate the journal on a self-sustaining independent basis is profoundly reassuring that its value has been clearly established. With the forthcoming transition to the private sector, the journal will be able to better serve the entire scientific community in the area of logistics research.

☆ U.S. GOVERNMENT PRINTING OFFICE : 1983 O - 407-572

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to Professor Herbert Solomon, Editor, Naval Research Logistics Quarterly, Department of Statistics, Sequoia Hall, Stanford University, Stanford, California 94305

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing will be expedited if four copies of the manuscript are submitted.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive reprints of their published papers.

For sale by the Superintendent of Documents, U.S. Government Printing Office Washington, D.C. 20402

NAVAL RESEARCH
LOGISTICS
QUARTERLY

DECEMBER 1982
VOL. 29, NO. 4
NAVSO P-1278

CONTENTS

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217