1994

A proposed method to estimate the effectiveness of delayed tasks

John Patrick Mullen
Iowa State University

Order Number 9424244

A proposed method to estimate the effectiveness of delayed tasks

Mullen, John Patrick, Ph.D.

Iowa State University, 1994

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

A proposed method to estimate the effectiveness of delayed tasks

by

John Patrick Mullen

A Dissertation Submitted to the

Graduate Faculty in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Department: Industrial and Manufacturing Systems Engineering
Major: Industrial Engineering

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Members of the Committee:

Signature was redacted for privacy.

Iowa State University
Ames, Iowa
1994

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

Overview

Command and control (C^2) is an essential part of critical ventures such as air traffic control, large-scale process management, disaster recovery, and military operations. Command and control communications and information (C^3I) systems are assemblages of equipment, people, and procedures that extend the range and effectiveness of C^2. Although modern C^3I systems are great improvements over those of the past, they do not fully utilize current technology and human potential. A major problem is that C^2 effectiveness predicted through C^3I analysis is often not replicated in the field [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].

This paper proposes that a major impediment to reliable C^3I design is suboptimization due to poor communications among the specialties involved. It presents a method that reduces such suboptimization. Building on previous work to improve unity [11, 12, 13], the paper presents the *focused measure of performance* (FMOP) as a means to more closely link C^2 effectiveness to C^3I performance and *pseudo-confidence intervals* (p.c.i.) as a way to assure that the apparent precision of C^3I analysis will not exceed that of the C^2 problem. It develops three sets of CMOPs and demonstrates how they may be used in numerous C^3I situations. It also presents a method that uses FMOPs and a Taguchi-like approach to seek profitable areas of

analysis in the C^2 and C^3I domains.

The emphasis in this paper is on time effectiveness. Examples of offered loads and effectiveness specifications are based on the Joint Services Communications Data Base (CDB) [14]. The remainder of this chapter outlines relevant problems in C^3I analysis and summarizes past work to overcome them. It concludes by outlining the aspects of the problem addressed in the remainder of the paper. Chapter 2 presents the proposed method together with evidence of its suitability to simulation C^3I analysis. Chapter 3 explores the robustness of the *hybrid distribution*, a key time distribution model used in Chapter 2. It establishes asymptotic and other bounds on estimation bias, showing the hybrid FMOP is more robust than others based on mean time or two-parameter distribution models. Finally, Chapter 4 summarizes the report, draws conclusions, and suggests directions for further research. Three appendices detail 1) conversion of CDB data, 2) the nature of autocorrelation, and 3) fine points of an extended example in Chapter 2.

Background on C^3I Systems

The need for C^3I systems

Morris defines command and control (C^2) as an interaction between two primary elements: the commander and the controlled unit [7]. Figure 1.1 shows how the C^3I system supports C^2 by transmitting and reformatting data, as needed, to provide effective communications between the command and the controlled unit.

Since controlled units are often far away from their commander, a telecommunications system is needed to convey messages in both directions. If there are many

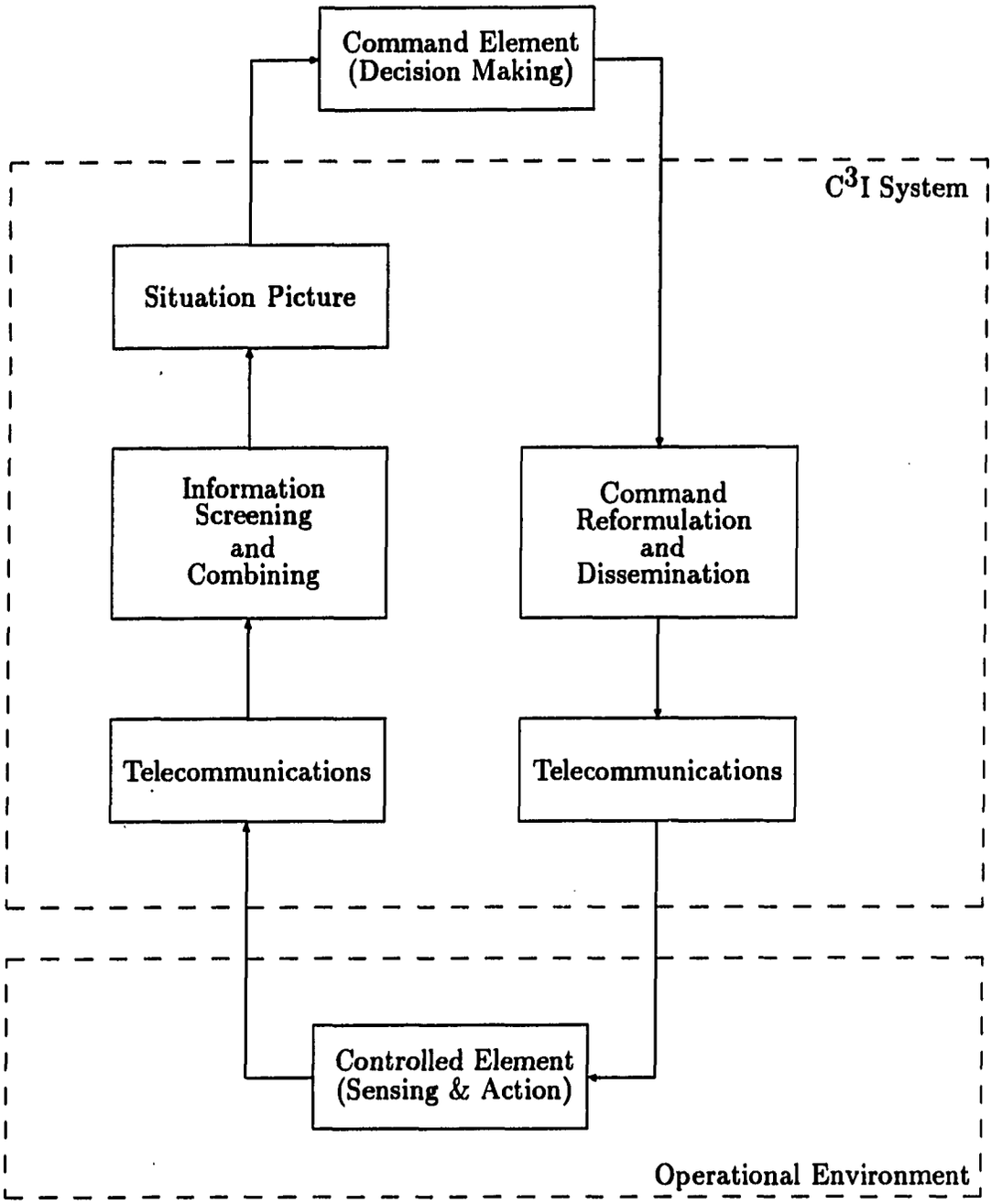


Figure 1.1: The Interrelationships of C^2 and C^3I

data sources, the commander will also need ancillary units at the command post to filter and organize incoming data, presenting a *situation picture*. For example, an air-traffic controller's display shows a composite representation of data from radars, IFF receivers, and tracking computers in an immediately-recognizable form. If there are many controlled units, the commander will need other ancillary units to decompose global commands into instructions for individual units. For example, if a fire chief orders fire-fighters to withdraw from a ridge, the staff would translate this command into individual orders for the several fire-fighting units in the vicinity. These ancillary units taken together with the communications subsystems form the C^3I system; a communications and information system that supports C^2.

The complexity of C^3I design

Although we posses the technology to build very effective C^3I systems, we have not yet done so. In 1985, White stated that this was due to informational chaos, the technological revolution, organizational chaos, and a lack of C^2 theory [8]. The situation has improved since, but not to the point of automating C^3I system design. In 1989, White assessed the C^3I problem in this way.

> Failures in C3 design process are, in part, evidence of too strong a faith in the symbolic side of systems analysis. We know it works, for we have seen it work. We know it doesn't work, for we have seen it fail. All we need now is the insight to anticipate both cases [10].

Others agree that the situation is better, but still mixed [15, 16, 17, 18, 19, 20].

One major problem is the need for diverse expertise. Command and control is partially a management problem and partially a leadership problem. Understanding

of command and control involves operations research, military expertise, ergonomics, human psychology, etc. Understanding of C^3I, on the other hand, involves electronics, wave-propagation theory, cybernetics, etc. There is very little overlap of expertise in these fields, yet members of all specialties must cooperate to produce an effective C^3I design. Other problems include the state of problem descriptions, current level of knowledge, and the lack of a uniform approach.

Foundations of C^3I Analysis

MCES: modularization of the problem

One important approach to the diverse expertise problem is the *Modular Command and Control Evaluation Structure* (MCES) [12]. At the 1984 Measures of Effectiveness for C3 Evaluation Symposium, AFG Eaglet challenged the conferees to derive a plan to measure C^3 effectiveness. The upshoot of this challenge was the MCES. This structure does not actually specify measures of performance, but evaluates them instead. Since then others have continued to develop and use the MCES [17, 19].

One outstanding feature of the MCES is concept of system boundaries and an associated hierarchy of measures:

- Measures of Force Effectiveness (MOFE)
- Measures of Effectiveness (MOE)
- Measures of Performance (MOP)
- Dimensional Parameters (DP).

Dimensional parameters are static features of the C^3I system such as weight, resolution, and baud rate. Demands placed upon the C^3I system do not affect them. *Measures of performance* are quantifications of C^3I system behavior, such as transportability, reliability, and throughput. This behavior depends upon DPs, the working environment, and demands of the supported C^2 system. *Measures of effectiveness* such as probability of detection, reaction time, and kill probability, quantify the ability of the C^2 system to perform its tasks. *Measures of force effectiveness* describe how well the entire unit performs. An example would be the probability of taking and holding a particular geographical location.

The analyst's ultimate goal is to determine MOFEs. Although good methods exist to do this, they are very costly. The intermediate problem, therefore, is to find good candidate C^3I systems using MOEs. The MCES concept, depicted in Figure 1.2, is to define MOPs so that each MOE is a function of the MOPs. This allows C^2 and C^3I experts to coordinate their independent efforts while isolating the problem of estimating MOPs from that of estimating effectiveness.

The key to this method is the MOP set and the MOP/MOE functions. Unless one can identify a measurable set of MOPs (θ) and an acceptable MOP/MOE function ($\mathcal{F}$), one cannot set up the necessary interface between C^2 and C^3I. Currently, there is no universal agreement on the exact nature of these MOP/MOE functions. If one were to estimate effectiveness values directly, one would have to repeat most of the work to consider other models of effectiveness. Using the MCES approach, one can consider one set of MOPs under several MOP/MOE functions, saving much of the simulation and analysis costs. (See Figure 1.2).

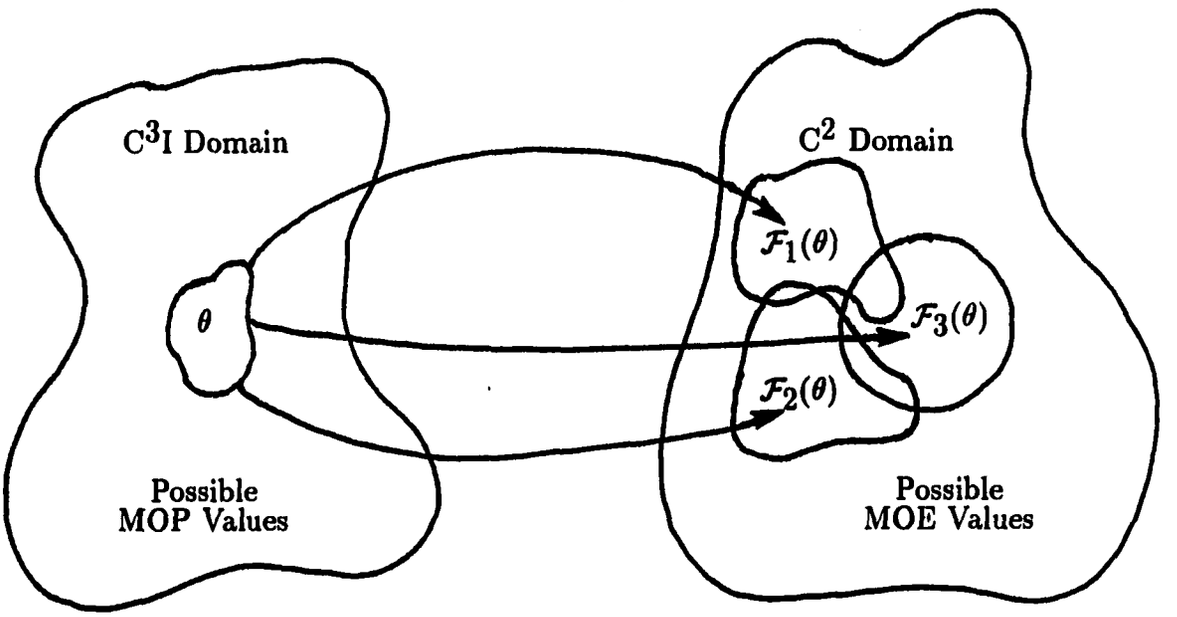


Figure 1.2: MOEs as Functions of MOPs

CDB: communications requirements specifications

One cannot proceed beyond dimensional parameters without first defining C^2 demands. Towards this end, the U.S. Army Signal Center at Ft. Gordon, GA developed and now maintains the *Communications Data Base* (CDB) [14]. The CDB details the communications needs of more than 300,000 communicator pairs. It specifies these needs independently of C^3I capabilities. All proponents of the U.S. Army Training and Doctrine Command have verified the CDB and the Combined Arms Center has validated it. The data base now contains only unclassified records, so developers may use it without compromising military security. The Signal Center continually reviews the CDB and issues annual updates. It is the only sanctioned description of battlefield communications needs for the U.S. Army.

Communications needs Command and control is a complex process. In addition, a command and control system consists of many interrelated C^2 subsystems. Beam classifies these subsystems in [21, ch. 6] by 1) the nature of the mission, 2) the organization's structure, 3) interdependence with other subsystems, and 4) information flow. Although this is far from an all-encompassing model of C^2, it does serve to describe C^2 requirements and probable communications flow.

Information flow patterns are derived from the other three classifications. The mission determines what activities are necessary. The organization's C^2 structure will dictate who will communicate with whom. Communications needs among the different C^2 subsystems arise from their interdependence. The volume of traffic will depend upon the environment and enemy force's actions as well. For example, a unit on a reconnaissance mission might normally maintain radio silence and report to its intelligence officer upon return. However, if it is trapped and in danger from the enemy, the unit leader might call to relay crucial information.

The CDB describes the communications needs in terms of mission, organizational unit both between and within military units. The unit of specification is the *needline*, a specific origin–destination requirement under specific conditions, and in C^3I-system-independent form. The CDB describes only needlines judged at least essential to C^2 activities. This limitation will remain since it focuses on requirements that most affect C^3I performance under battlefield conditions.

Effectiveness measures In addition to the load specification, the CDB contains information for effectiveness rating. The "Perishability" code ($\mathcal{P}$ in Table 1.1) indicates the acceptable ranges of completion times for each needline. The

Table 1.1: Communications Data Base Perishability Codes [14, p. 9]

Perishability Code ($\mathcal{P}$)	Time Range	Perishability Code ($\mathcal{P}$)	Time Range
0	> 8 HRS	6	1-10 MIN
1	4-8 HRS	7	25-59 SEC
2	3-4 HRS	8	11-24 SEC
3	2-3 HRS	9	5-10 SEC
4	1-2 HRS	A	1-4 SEC
5	10-60 MIN	B	< 1 SEC

Table 1.2: CDB Cost of Failure Codes [14, p. 2]

Code	Description
I	Indispensable
C	Critical
E	Essential

"Cost of Failure" code (Table 1.2) reflects the ability of a commander to complete the specified mission, if he or she does not receive that particular message in time. The "Purpose" and "Function" codes define the mission. However, because there is no standard way to assign numerical values to the "Cost of Failure" codes, the CDB does not precisely quantify the impact of the needline on the mission.

Summary The CDB is a source of C^2 requirements that is widely accepted and available. It describes C^3I–independent loads and partially specifies effectiveness measures. Since it is understandable to both C^2 and C^3I experts, it acts as an interface between the two groups. The CDB will be the basis of all offered system loads and MOP-MOE conversion functions in this paper.

Timeliness: the effect of delay on value

To some degree, response time always affects task effectiveness. Raatikainen claims that given the current state of technology, the response time has become "the most important single performance indicator" [22, p. 190]. Certainly, faster is better, but system designers must allocate limited resources to achieve the best possible overall effectiveness. Thus, they must ask "How much better?"

Let $\mathcal{V}(t)$ represent the residual effectiveness of a completed task as a function of its execution time, t. Then, given a representative set of n tasks, an obvious measure of system effectiveness would be the mean value of executed tasks

$$\mathrm{E}[\mathcal{V}(T)] = \sum_{j=1}^{m} \mathrm{E}\left[\mathcal{V}_j(T_j)\right] \tag{1.1}$$

where $\mathcal{V}_j(t)$ is the residual effectiveness function and T_j the random completion time of Task i. Letting $I_j = \mathcal{V}(0)$ and $\mathcal{T}_j(t) = \mathcal{V}_j(t)/I_j$ leads to

$$\mathrm{E}[\mathcal{V}(T)] = \sum_{j=1}^{m} I_j \, \mathrm{E}\left[\mathcal{T}_j(T_j)\right] \tag{1.2}$$

where I_j is the *importance* of Task i, and[1]

$$\mathrm{E}[\mathcal{T}_j(T_j)] = \int_0^\infty \mathcal{T}_j(t) d\mathrm{F}_j(t) \tag{1.3}$$

is the mean time effectiveness or *timeliness* for that task. Equation (1.2) clearly shows the logical separation of C^2 and C^3I functions. Task descriptions, I_j, and $\mathcal{T}_j(t)$ are determined in the C^2 arena without considering the C^3I system. $\mathrm{F}_j(t)$ is determined by C^3I experts using specific C^3I models. (1.3) shows a key interface

[1]The Stieltjes integral is used in this paper because, unlike the Riemann integral, it is defined for functions that are not continuous [23, ch. 7].

point. C^3I experts cannot determine $E[\mathcal{T}_j(T_j)]$ without descriptions of the C^2 load and timeliness, while C^2 experts cannot determine it without accurate C^3I models.

Properties of timeliness Cothier and Levis defined *timeliness* to be "a systems ability to respond within an allotted time [24, p. 844]." Extending this idea, the C3INAT team defined it to be the mean effect of completion time upon task effectiveness. The timeliness function, $\mathcal{T}(t)$, predicts the effect of delay on the value of a task completed at time t. Using very minor assumptions, one can show four properties of $\mathcal{T}(t)$. By its definition,

$$\mathcal{T}(0) = 1.0. \tag{1.4}$$

Assuming delay never improves value, $\mathcal{T}(t)$ is a non-increasing function of t and

$$\frac{d\mathcal{T}(t)}{dt} \leq 0 \tag{1.5}$$

whenever the derivative exists. Ignoring the possibility of negative value, it is reasonable to assume that eventually the task will become worthless, or

$$\lim_{t \to \infty} \mathcal{T}(t) = 0. \tag{1.6}$$

Any right-continuous non-increasing function meeting these conditions can represent timeliness. Finally, because $\mathcal{T}(t)$ is bounded, (1.4) through (1.6) imply it cannot be a linear function of time. Beyond these statements, the exact nature of timeliness will depend upon the C^2 requirements for the task.

Problems in C^{3}I Analysis

The following describes the specific problems which this paper addresses.

Inadequacy of mean time to predict E[$\mathcal{T}(T)$]

Because one can get accurate, robust estimators of E[T] in most situations, it is a common practice is to rank systems on the basis of mean time. The following example illustrates the inadequacy of this method to estimate E[$\mathcal{T}(T)$].

Example 1.1 *An analyst must choose between two communication links. The first is a single hop with a mean transmission rate of two messages per minute. The second is a two-hop link with each hop having a mean transmission rate of three messages per minute. Message initiations follow a Poisson distribution with a mean of one message per minute and transmission times are exponentially distributed. Furthermore, transmission times on the two series hops are independent and the buffers in both links are very large.*

The timeliness model is the two-parameter exponential model,

$$\mathcal{T}_2\left(t;\tilde{t},t_0\right) = \begin{cases} 1.0 & 0 \le t < t_0 \\ \exp\left(\dfrac{t-t_0}{\tilde{t}}\right) & t \ge t_0. \end{cases} \tag{1.7}$$

Here t_0 is the time of initial decay and $\tilde{t}$ is the mean of the exponential decay for times greater than t_0. The values of t_0 and $\tilde{t}$, which differ by $\mathcal{P}$ class[2]*, are listed in Table 1.3.*

[2]The parameters are chosen so that $\mathcal{T}_2(t_\ell) = 0.9$ and $\mathcal{T}_2(t_h) = 0.1$. Full details of their determination are in Appendix A.

Table 1.3: Reference Times for $\mathcal{T}_2(t;\tilde{t},t_0)$ with $\mathcal{T}_\ell = 0.9$ and $\mathcal{T}_h = 0.1$

Perishability ($\mathcal{P}$)	Units	t_ℓ	t_h	t_0	$\tilde{t}$
0	Hrs.	8	24	7.233	7.282
1	Hrs.	4	8	3.808	1.820
2	Hrs.	3	4	2.952	0.455
3	Hrs.	2	3	1.952	0.455
4	Hrs.	1	2	0.952	0.455
5	Min.	10	60	7.602	22.756
6	Min.	1	10	0.568	4.096
7	Sec.	25	60	23.322	15.929
8	Sec.	11	25	10.329	6.372
9	Sec.	5	11	4.712	2.731
A	Sec.	1	5	0.808	1.820
B	Sec.	0.046	1	0.000	0.434

Mean time A commonly-studied MOP for time is E[T], the mean time. The one-hop link is a (M/M/1) queue with $\lambda = 1$ and $\mu_1 = 2$, so E[T] = 1 minute. The first hop in the two-hop link is also a (M/M/1) queue, but with $\lambda = 1$ and $\mu_2 = 3$. Burke's theorem [25] proves that the output of the first hop is a Poisson process with mean λ, so the second hop is also a (M/M/1) queue. Therefore, E[T] = $2/(\mu_2 - \lambda) = 1.0$ minute, and E[T] is the same for both links.

Mean timeliness Effectiveness is a function of timeliness. From (1.3) and (1.7), the mean timeliness for the two-parameter exponential function is

$$\mathrm{E}\left[\mathcal{T}_2\left(T;\tilde{t},t_0\right)\right] \;=\; \mathrm{F}\left(t_0\right) + \int_{t_0}^{\infty} \exp\left(\frac{t-t_0}{\tilde{t}}\right) d\mathrm{F}(t). \tag{1.8}$$

Because completion time in the one-hop link is exponentially-distributed, with mean $\theta_1 = 1/(\mu_1 - \lambda)$,

$$\mathrm{E}\left[\mathcal{T}_2\left(T;\tilde{t},t_0\right)|\theta_1\right] \;=\; 1 - \left(\frac{\theta_1}{\theta_1+\tilde{t}}\right)\exp\left(\frac{-t_0}{\theta_1}\right). \tag{1.9}$$

Table 1.4: Mean Timeliness for Both Links in Example 1.1

$\mathcal{P}$	$\mathrm{E}[\mathcal{T}_2(T;\mathcal{P},\mathcal{T}_\ell,\mathcal{T}_h)]$, with $\mathcal{T}_\ell = 0.9$ and $\mathcal{T}_h = 0.1$	
Code	One-Hop Link	Two-Hop Link
5	1.000	1.000
6	0.889	0.894
7	0.464	0.362
8	0.239	0.112
9	0.116	0.028
A	0.042	0.005
B	0.007	0.000

Because the distribution of T on the two-hop link is Erlang-2 [26, p. 124] with mean $\theta_2 = 2/(\mu_2 - \lambda)$,

$$\mathrm{E}[\mathcal{T}_2(T;\tilde{t},t_0)|\theta_2] = 1 - \left(1 + \frac{2t_0}{\theta_2}\right)\exp\left(-\frac{2t_0}{\theta_2}\right) + \left(\frac{2\tilde{t}}{\theta_2 + 2\tilde{t}}\right)^2\left[1 + \left(\frac{2}{\theta_2} + \frac{1}{\tilde{t}}\right)t_0\right]\exp\left(-\frac{2t_0}{\theta_2}\right). \quad (1.10)$$

Table 1.4 lists mean timeliness values by message class, for the two links.

Analysis Although the two links have the same mean completion times, they do not have the same mean timeliness. In addition, the relative dominance depends on the timeliness function. The two-hop link is as good or better if $\mathcal{P} \leq 6$ for all messages, but worse if $\mathcal{P} \geq 7$. Thus, the analyst's preference for one or the other link will depend upon the mix of needlines and their relative importance. Since $\mathrm{E}[T]$ is the same in both links above, no function of $\mathrm{E}[T]$ alone can reflect their differences in timeliness.

Inadequacy of two-parameter distribution models

Since the shape of F(t) is a function of its variance, a two-parameter model of F(t) should lead to some improvement over using E[T] alone. Example 1.2 describes a more general case than Example 1.1 that will illustrate the degree of improvement as well as some remaining problems.

Example 1.2 *The message processing time in a complex system follows a gamma distribution with mean θ and standard deviation σ. Although the analyst can estimate θ and σ closely, he or she does not know T is a gamma random variable. As before, the timeliness model is the two-parameter exponential model $\mathcal{T}_2(t;\tilde{t},t_0)$ and the analyst wishes to compute $E[\mathcal{T}_2(T)]$.*

Part a of Figure 1.3 illustrates the difference between the true mean timeliness and estimated timeliness[3] in Example 1.2, assuming the analyst characterizes F(t) as exponential with mean θ. This error is shown as the exponential estimate less the true value as a function of θ and the coefficient of variation φ[4] Because the exponential distribution is based only on θ, it is logically equivalent to estimating E[$\mathcal{T}(T)$] on the basis of E[T] alone. As expected, the possible error is huge, with a width of nearly one, the full range of E[$\mathcal{T}(T)$]. Part b of this figure shows the error that exists if the analyst assumes a log-normal distribution. Certainly, the lognormal distribution is a great improvement over the exponential approximation, although it still can have a large error when $\varphi > 1$.

[3]The details of these computations are in Chapter 3.

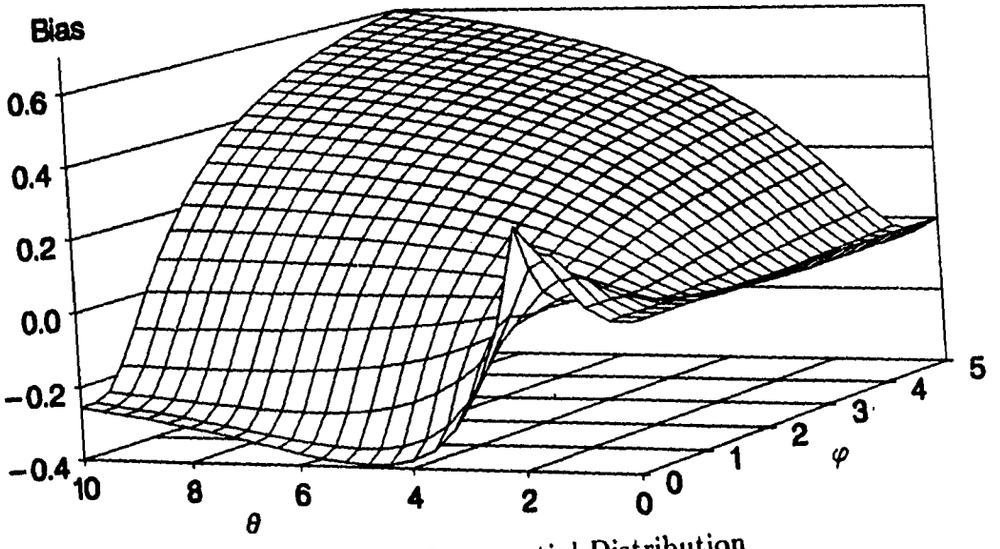
[4]The coefficient of variation is used, rather than σ, since it represents shape, independent of scale.

When φ is small, all distributions of non-negative numbers are similar, but when $\varphi > 1$, differences between the true distribution, $F_T(t)$ and any two-parameter model can affect $E[\mathcal{T}(T)]$ significantly. Since one is practically limited to the first two moments of distributions in simulation, the potential error can be a serious problem, especially when variance is relatively high.

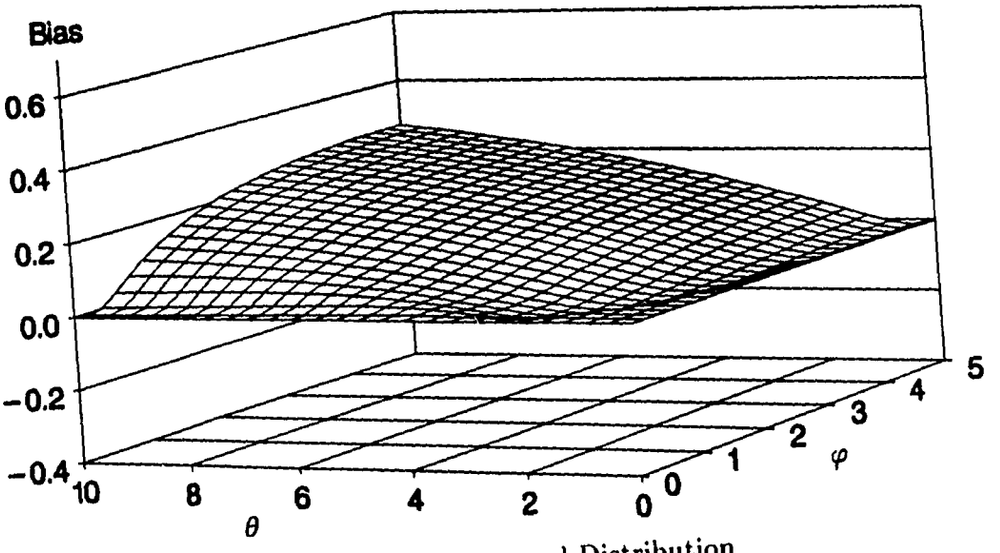
Autocorrelation

Economic constraints require users of C^3I systems to share resources. Thus, queues will form from time to time, especially when the system is heavily loaded. This leads to completion times that are self- or *autocorrelated.* Autocorrelation, discussed in Appendix B, 1) causes initial conditions to persist, 2) causes seemingly independent runs to be correlated, and 3) frustrates efforts to estimate variance directly. The first two effects can either be viewed as part of the analysis or reduced in significance by sufficiently long warm-up times. Because $E[X + Y] = E[X] + E[Y]$ even if X and Y are correlated, the sample mean will be an unbiased estimator of the process mean as long as the biased observations from the warm-up period are discarded. However, autocorrelation will always affect estimates of higher moments and attempts to determine the shape of T's distribution..

Because successive completion times are positively correlated, the sample variance of the output trace is less than $\mathrm{Var}(T)$. In addition, since its effect occurs most at congested times, it distorts the shape of $F(t)$. This, serial correlation renders useless most traditional statistical analysis methods. Simulation studies of autocorrelated systems deal with means, percentages, or percentiles [27]. In general, it is very difficult to determine the true distribution of T. In addition, because one cannot



a. Using the Exponential Distribution



b. Using the Lognormal Distribution

Figure 1.3: Error Estimating $E[T_2(T;1,2)]$, using the Indicated Distribution when T Follows a Gamma Distribution having Mean θ and Coefficient of Variation φ

estimate φ directly, the bias in estimating $E[\mathcal{T}(T)]$ may be much greater than that shown above.

Imprecision of C^2 data

The CDB is a remarkable collection of information. However, although it has been checked by numerous experts, its entries have no definite range of precision. Because no parameter value can be known exactly, each reviewer assumed some degree of precision. Since this is not stated either in the documentation or in the database itself, one has no idea how precise the entries are. In addition, there is no statement of assumed distributions. Some needlines seem to be regularly scheduled and others would seem to occur more randomly, but no mention of variance is made in the database. Because the arrival and message-length distributions can greatly affect the distribution of time, one must speculate on their range to get accurate results.

Current C^3I analysis assumes the C^2 data is exact. Although C^3I analysts often perturb the C^3I model to reflect changes in the load due to changes in the scenario, they usually make no allowance for the range of precision in mean arrival rates, message length, etc. The following example illustrates this problem.

Example 1.3 *The analyst must choose between the two links of Example 1.1. The only design parameter is the index of the system. There are no environmental factors. The transmission rates, μ_1 and μ_2 are exactly known, so there are no internal noise factors. However, λ is only known to be within the interval $(0.5, 1.5)$. Finally, the analyst will again use the two-parameter exponential timeliness model.*

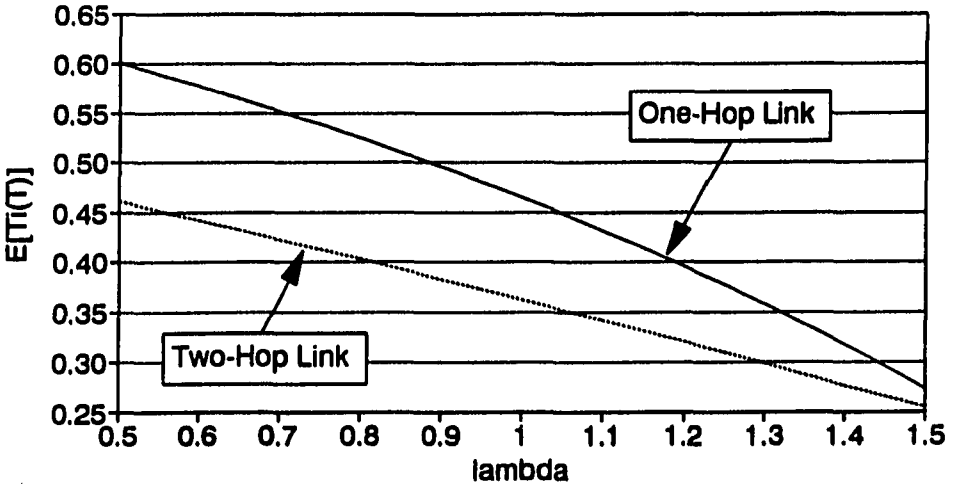


Figure 1.4: $\mathrm{E}[\mathcal{T}(T)]$ as a Function of λ in Example 1.3 at $\mathcal{P} = 7$

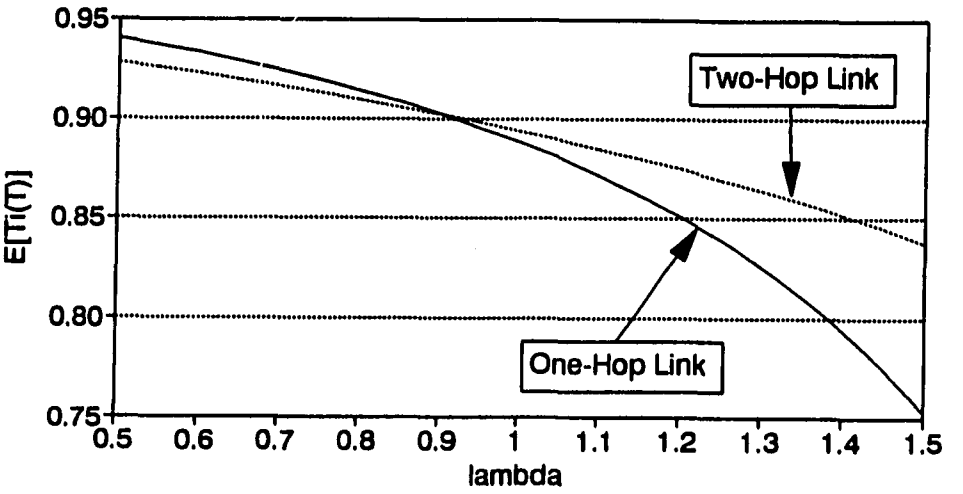


Figure 1.5: $\mathrm{E}[\mathcal{T}(T)]$ as a Function of λ in Example 1.3 at $\mathcal{P} = 6$

Figure 1.4 shows a plot of $E[\mathcal{T}(T)]$ as a function of λ, assuming $\mathcal{P} = 7$. In this case, the range of λ causes no concern, since Link 1 is superior to Link 2 over the entire interval. However, Figure 1.5 shows that when $\mathcal{P} = 6$, the one-hop link is better when λ is small and the two-hop when λ is large. This range of λ would lead to significant differences in effectiveness. The actual situation could be worse. For example, if there were two message classes, one with $\mathcal{P}_1 = 6$ and another with $\mathcal{P}_2 = 7$, one could specify mixtures of these classes in which either link is superior for any $\lambda \in (1.0, 1.5)$.

There are two serious consequences of this unknown C^2 data precision. First of all, the C^3I analyst may not exercise the C^3I model over enough of the potential C^2 load to discover system weaknesses. For example, in Figure 1.5, if the analyst simply looked at $\lambda < 0.9$, he or she might conclude that the one-link option was better. Secondly, simulation results are often reported as confidence intervals. Again, in Figure 1.5, if the analyst examined results for the two links at $\lambda = 0.8$ with a confidence interval width of ± 0.05, he or she might execute the simulation models many times in order to show statistically the one-link system is superior. However, if the analyst exercised the models for $\lambda \in [0.5, 1.5)$, he or she would know that further simulation would be pointless.

Overview of Work Presented in this Dissertation

The working assumption is that deficiencies in C^3I analysis are primarily due to suboptimization arising from poor communication between the C^2 and C^3I communities. The writer proposes to improve this communication by

1. increasing the degree of communication within the MOP-MOE function,

2. realistically considering the precision of the C^2 data and MOE functions, and

3. presenting a scheme to identify those regions of the C^2 and C^3I domain that need further refinement.

Additionally, the proposed procedure is economical, robust, and usable in the presence of autocorrelation.

Chapter two presents the proposed method and its development. Chapter three examines the robustness of the estimators used in the method. Finally, Chapter four summarizes the work, draws conclusions, and outlines directions for further research.

CHAPTER 2. THE PROPOSED METHOD

Introduction

This chapter presents two analytical tools that improve the reliability of the C^3I modeling process by enhancing C^2-C^3I communications. The first tool is the *focused measure of performance* (FMOP), an extension of the MCES concept that more closely binds the C^2 and C^3I communities. The second, the *pseudo-confidence interval* p.c.i., is used with an extension of Taguchi's method to translate C^2 modeling and evaluation uncertainty into intervals of C^3I effectiveness. The end result is an extension of both the MCES and Taguchi's method that is a more reasonable predictor of C^3I system effectiveness. The discussion centers on system-wide measures of time effectiveness. It assumes the situation is complex enough to warrant simulation. It also presumes a sufficiently-defined set of C^2 specifications and effectiveness measures, similar to that found in the CDB.

The following first presents FMOPs and details their use with exact models and independent trails. It then introduces the p.c.i. and an extension of the Taguchi method as a means of dealing with inexact models. Next, it suggests an approach that guides effort within the joint C^2-C^3I venture. Finally, the discussion turns to the special problems associated with the steady-state analysis in the presence of autocorrelation.

Measures of System Time Effectiveness

Naturally, necessary FMOPs (and MOPs) for analysis are dictated by MOE functions. The following assumes some variation of one of the following two measures of system time effectiveness is the MOE of concern. As noted below, the mean timeliness for needlines of concern is a sufficient FMOP for either.

Mean system time effectiveness

Suppose one has accurate effectiveness functions for all significant traffic on a C^3I system. Then, one logical measure of performance would be the mean system effectiveness:

$$\mathrm{E_S}[\mathcal{V}(T)] \triangleq \sum_{j=1}^{m} \left(\frac{\lambda_j}{\lambda}\right) \mathrm{E}[\mathcal{V}_j(T_j)] \tag{2.1}$$

where λ_j is the arrival rate, $\mathcal{V}_j(t)$ the time effectiveness function, and T_j the random completion time for the j-th needline. Here, $\lambda = \sum_{j=1}^{m} \lambda_j$ is the over-all arrival rate among the needlines of concern, which may be less than the total system arrival rate. Using the timeliness function, one can express (2.1) as:

$$\mathrm{E_S}[\mathcal{V}(T)] = \sum_{j=1}^{m} \left(\frac{\lambda_j I_j}{\lambda}\right) \mathrm{E}[\mathcal{T}_j(T_j)] \tag{2.2}$$

where I_j is the importance factor and $\mathcal{T}_j(t)$ is the timeliness function for Needline j. Thus, one may define *mean system timeliness* to be the normalized mean time effectiveness, or

$$\mathrm{E_S}[\mathcal{T}(T)] \triangleq \frac{\mathrm{E_S}[\mathcal{V}(T)]}{\mathrm{E_S}[\mathcal{V}(0)]}$$

$$= \frac{\sum_{j=1}^{m} \lambda_j I_j \mathrm{E}[\mathcal{T}_j(T_j)]}{\sum_{j=1}^{m} \lambda_j I_j}. \tag{2.3}$$

This measure indicates the time effectiveness of the system on a scale from zero to one. C^2 analysts can specify minimum system performance as single value, E_{MIN}. Because (2.3) depends upon the ratios of importance levels, it is sensitive to their relative importance, rather than their values.

Minimum system timeliness

It may be that although C^2 experts can identify importance classes, they cannot precisely quantify a numerical relationship among them. In this case, a reasonable measure of performance would be *minimum system timeliness*:

$$\min_S[\mathcal{T}(T)\,|I] \triangleq \min_{j \ni I_k = I} \mathrm{E}[\mathcal{T}_j(T_j)]. \tag{2.4}$$

Here, one estimates timeliness by importance class, and minimum specifications would be given by class.

Usefulness of needline mean timeliness

In either case above, given C^2 specifications and C^3I estimates of individual needline mean timeliness values, one may estimate $E_S[\mathcal{T}(T)]$ through (2.3) and $\min_S[\mathcal{T}(T)\,|I]$ through (2.4) . The discussion now turns to estimates of individual needline mean timeliness values. To simplify notation, the subscript j will be dropped, unless more than one needline is considered simultaneously.

Needline Mean Timeliness as a FMOP

The FMOP concept

The MCES concept makes C^3I analysis more efficient by segregating tasks by specialty and using parameters to make consideration of multiple MOE functions more efficient. However, traditional MOPs do not convey enough information between C^2 experts and C^3I analysts. As illustrated in Figure 2.1, the *focused measure of performance* (FMOP) improves the degree of information flow, making it more useful in estimating effectiveness. Like a MOP, it is an observable system characteristic related to the quantity of interest. However, unlike a MOP, the FMOP also includes a region of interest (Ω). For example, mean bit error rate is a MOP. The probability that there are more than 10^6 bits between errors is a FMOP with $\Omega = [10^6, \infty)$. In the case of the CDB and timeliness, each $\mathcal{P}$ code identifies a distinct $\Omega = (t_\ell, t_h)$ as the interval of highest interest. A delay less than t_ℓ has little effect on that task's effectiveness. A delay greater than t_h will render the task's completion nearly worthless.

In theory, measures of performance are system characteristics that are independent of the context. In practice, they are not, because they are estimated by exercising a C^3I model under a specific set of offered loads in a specific series of scenarios. One loses little by specifying the range of interest in C^3I performance, as long as one does so in terms that are independent of the C^3I system. Because C^2 experts specify the FMOP's region of focus and C^3I experts determine the FMOP's value, it is a more complete form of communication than a MOP between these two groups. The error FMOP above tells C^2 analysts they can ignore errors in a system with a mean bit

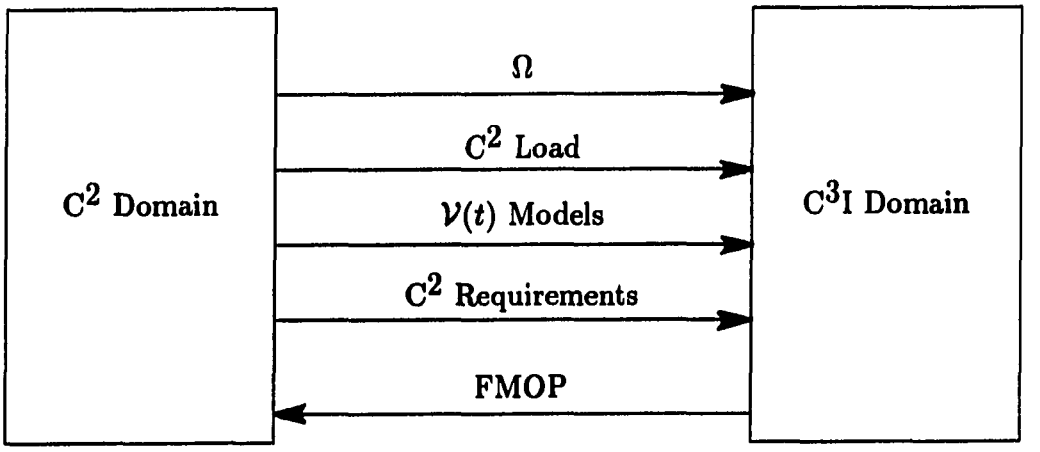


Figure 2.1: How the FMOP Enhances Information Flow

error rate of 10^{-12} and need not consider a system with one of 10^{-3}.

C^3I analysts can use the FMOP to predict effectiveness better than conventional MOPs. For example, if $\mathcal{T}_\ell \triangleq \mathcal{T}(t_\ell) = 0.9$ and $\mathcal{T}_h \triangleq \mathcal{T}(t_h) = 0.1$, then

$$\begin{aligned} \mathrm{E}[\mathcal{T}(T)] &= \int_0^{t_\ell} \mathcal{T}(t)d\mathrm{F}(t) + \int_{t_\ell}^{t_h} \mathcal{T}(t)d\mathrm{F}(t) + \int_{t_h}^{\infty} \mathcal{T}(t)d\mathrm{F}(t) \\ &\approx 0.1 + \int_{t_\ell}^{t_h} \mathcal{T}(t)d\mathrm{F}(t). \end{aligned} \tag{2.5}$$

The approximation (2.5) differs from true mean timeliness by no more than 0.1. Thus, an analyst would need to estimate how $\mathcal{T}$(t) and F(t) behave on Ω far more closely than in their tails. This is fortunate, since one usually knows little about the tail behavior of these functions. Because analysts do not know the region of greatest C^2 impact with unfocused MOPs, they may allow the behavior in the tails of the timeliness function to dominate the MOP estimate, leading to an arbitrarily large error in the effectiveness estimate.

Characteristics of mean timeliness

For a single needline, mean timeliness is the mean value of the timeliness function:

$$\mathrm{E}[\mathcal{T}(T)] \;=\; \int_0^\infty \mathcal{T}(t)d\mathrm{F}_T(t). \tag{2.6}$$

Using integration by parts and the fact that both functions are bounded, we get

$$\mathrm{E}[\mathcal{T}(T)] \;=\; -\int_0^\infty \mathrm{F}_T(t)d\mathcal{T}(t). \tag{2.7}$$

The composite function $\mathrm{F}_\mathcal{T}(\tau) \triangleq \mathrm{F}_T\left[\mathcal{T}^{-1}(\tau)\right]$ may not be well-defined, since $\mathcal{T}(t)$ is not necessarily one-to-one. However, the function

$$\mathcal{T}^{-1^\star}(\tau) \;\triangleq\; \max_{t \ni \mathcal{T}(t)=\tau}(t) \tag{2.8}$$

is. Thus, given $\mathrm{F}_T(t)$ and $\mathcal{T}(t)$, one may define the probability distribution of timeliness, $\mathrm{F}_\mathcal{T}(\tau) = \mathrm{F}_\mathcal{T}\left[\mathcal{T}^{-1}(\tau)\right]$. Figure 2.2 illustrates typical plots of this function.

Equation (2.7) shows that the area under $\mathrm{F}_\mathcal{T}(\tau)$ in Figure 2.2 is

$$\mathrm{E}[\mathcal{T}(T)] \;=\; \int_0^1 d\,\mathrm{F}_\mathcal{T}(\tau). \tag{2.9}$$

This implies the timeliness function primarily determines the area of significance for evaluating $\mathrm{E}[\mathcal{T}(T)]$ and the distribution of T is important only in its ability to predict the distribution of $\mathcal{T}(\tau)$.

Given $\mathcal{T}_\ell$ and $\mathcal{T}_h$, but no other information about $\mathcal{T}(t)$, detailed information about $\mathrm{F}(t)$ is of no value. Theorem 3.5 shows that if $\mathcal{T}_a(t) \geq \mathcal{T}_b(t)$ for all T, then $\mathrm{E}\left[\mathcal{T}_a(T)\right] \geq \mathrm{E}\left[\mathcal{T}_b(T)\right]$. Therefore, from (2.6)

$$\begin{aligned}\mathrm{E}[\mathcal{T}(T)] \;&\leq\; \int_0^{t_\ell} d\mathrm{F}(t) + \mathcal{T}_\ell \int_{t_\ell}^{t_h} d\mathrm{F}(t) + \mathcal{T}_h \int_{t_h}^{\infty} d\mathrm{F}(t) \\ &=\; p_\ell + \mathcal{T}_\ell\left(p_h - p_\ell\right) + \mathcal{T}_h\left(1 - p_h\right)\end{aligned} \tag{2.10}$$

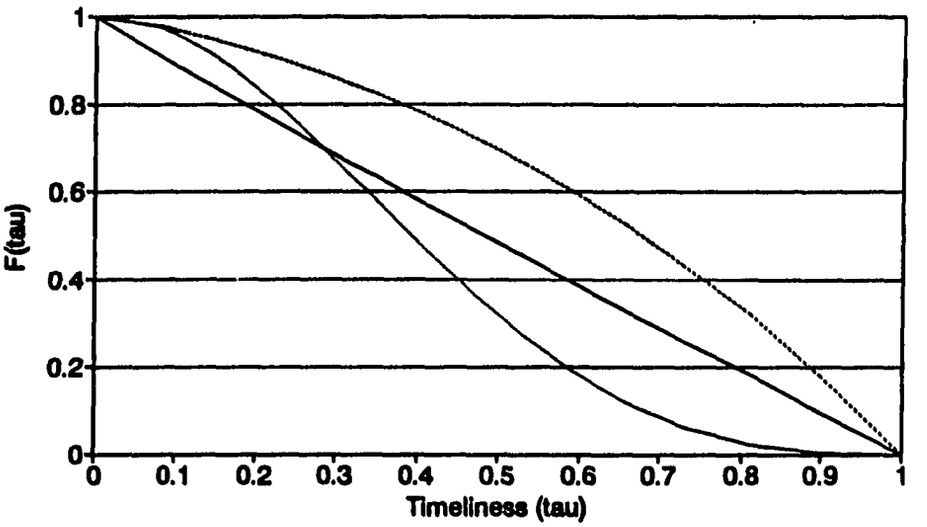


Figure 2.2: Sample Plots of the Probability Distribution of $\mathcal{T}(\tau)$

and

$$\begin{aligned} \mathrm{E}[\mathcal{T}(T)] &\geq \mathcal{T}_\ell \int_0^{t_\ell} d\mathrm{F}(t) + \mathcal{T} \int_{t_\ell}^{t_h} d\mathrm{F}(t) \\ &= \mathcal{T}_\ell p_\ell + \mathcal{T}_h \left(p_h - p_\ell\right). \end{aligned} \tag{2.11}$$

Unless one knows more about $\mathcal{T}(t)$, knowledge of the actual distribution of $\mathrm{F}(t)$, other than $\mathrm{F}(t_\ell) = p_\ell$ and $\mathrm{F}(t_h) = p_h$ cannot be used to determine tighter bounds on $\mathrm{E}[\mathcal{T}(T)]$.

Single-probability bounds on timeliness

Although bounding $\mathrm{E}[\mathcal{T}(T)]$ from a single probability estimate seems terribly crude, it is an effective way to screen many systems. One reason for this is the large differences among systems. A second is the relatively low data-collection and computational load. A third is the undiluted power of statistical tests based upon

single parameters. Additionally, because this estimate makes the fewest assumptions about $\mathcal{T}(t)$, it produces the most widely-reaching statements about $\mathrm{E}[\mathcal{T}(T)]$.

Probability estimators Suppose $\{T_i; i = 1, \ldots, n\}$ is a random sample from some time distribution with finite moments. Given a specific time, t_x, one may define $\{X_i; i = 1, \ldots, n\}$, where

$$X_i = \begin{cases} 1 & \text{if } T_i \leq t_x \\ 0 & \text{if } T_i > t_x. \end{cases} \tag{2.12}$$

Because the T_i are i.i.d., $N_x = \sum_{i=1}^{n} X_i$ is a binomial random variable and $\mathrm{E}[X] = p_x$, the probability that $T \leq t_x$. Furthermore, $\mathrm{Var}(X) = p_x(1-p_x)/n$. The estimate n_x/n, where n_x is the number of $T_i \leq t_x$, is an unbiased, maximum-likelihood estimator of p_x. In addition, $\widehat{\mathrm{Var}}(\hat{p}_x) = [\hat{p}_x(1-\hat{p}_x)]/n$. One may use this method to estimate any probability, including p_ℓ and p_h.

Bounds on mean timeliness From (2.7)

$$\begin{aligned} \mathrm{E}[\mathcal{T}(T)] &= -\int_0^{t_h} \mathrm{F}(t)d\mathcal{T}(t) - \int_{t_h}^{\infty} \mathrm{F}(t)d\mathcal{T}(t) \\ &\leq p_h\left(1-\mathcal{T}_h\right) + \mathcal{T}_h. \end{aligned} \tag{2.13}$$

Because (2.13) is a linear function of p_h, it follows that an upper one-sided $(1-\alpha) \times 100\%$ confidence limit for (2.13) is:

$$\hat{\mathrm{E}}^{+}_{1\alpha}\left[\mathcal{T}(T)\,\middle|\,\mathcal{T}_h, \hat{p}_h\right] = \left(1-\mathcal{T}_h\right)\hat{p}^{+}_{h\alpha} + \mathcal{T}_h \tag{2.14}$$

where $\hat{p}^{+}_{h\alpha}$ is the upper limit end of a one-sided $(1-\alpha) \times 100\%$ confidence interval for p_h. For example, using the normal approximation to the binomial,

$$\hat{\mathrm{E}}^{+}_{1\alpha}\left[\mathcal{T}(T)\,\middle|\,\mathcal{T}_h, \hat{p}_h, n\right] = \left(1-\mathcal{T}_h\right)\left[\hat{p}_h + |z_\alpha|\sqrt{\frac{\hat{p}_h(1-\hat{p}_h)}{n}}\right] + \mathcal{T}_h \tag{2.15}$$

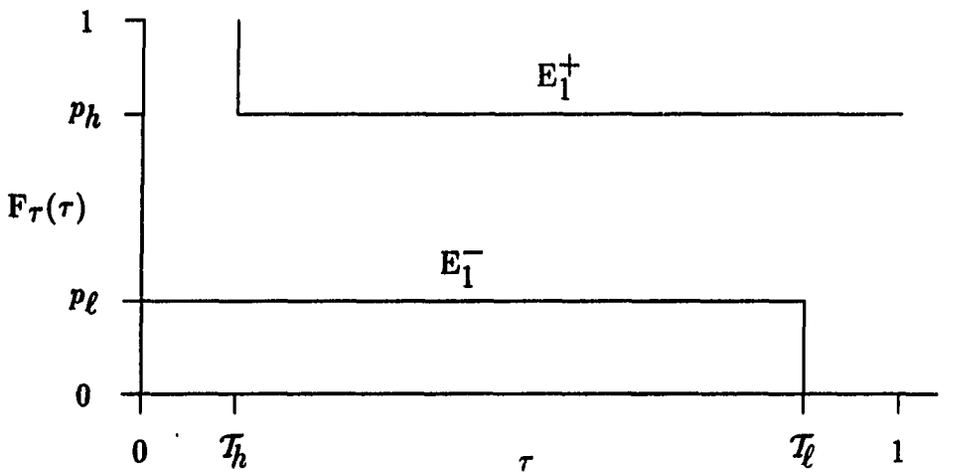


Figure 2.3: One-Estimate Bounds on E[$\mathcal{T}(T)$] for an Individual Needline

where z_α is the value of the standard normal variate at α.

Because $\hat{E}^+_{1_\alpha}[\cdot] < E_{MIN}$ is evidence that timeliness is unsatisfactory, this upper bound is an appropriate tool for eliminating poor systems. Similarly, one may define the lower bound

$$\hat{E}^-_{1_\alpha}\left[\mathcal{T}(T)\middle|\mathcal{T}_\ell, \hat{p}_\ell\right] = \mathcal{T}_\ell \hat{p}^-_{\ell_\alpha} \tag{2.16}$$

an approximate lower $(1-\alpha) \times 100\%$ c.l. of

$$\hat{E}^-_{1_\alpha}\left[\mathcal{T}(T)\middle|\mathcal{T}\ell, \hat{p}_\ell, n\right] = \mathcal{T}_\ell \left[\hat{p}_\ell - |z_\alpha| \sqrt{\frac{\hat{p}_\ell(1-\hat{p}_\ell)}{n}}\right] \tag{2.17}$$

to identify superior systems. These bounds are pictured in Figure 2.3

Two-probability bounds on mean timeliness

Joint probability estimators The confidence intervals above assume the analyst estimates only one of p_ℓ or p_h. If he or she uses both, the confidence intervals

are no longer exact. One could use a Bonferroni approximation, replacing z_α by $z_{(\alpha/2)}$ in (2.17) and (2.14) above, but this ignores the positive correlation of p_ℓ and p_h. Besides, the analyst is usually interested in just one of $\mathrm{E}^+[\mathcal{T}(T)]$ or $\mathrm{E}^-[\mathcal{T}(T)]$. Thus, the analyst would want two-probability bounds, rather than simultaneous one-probability bounds.

Again, assume $\{T_i; i = 1, \ldots, n\}$ is a random sample from some time distribution with finite moments. Let

$$X_{1i} = \begin{cases} 1 & \text{if } T_i \leq t_\ell \\ 0 & \text{if } T_i > t_\ell \end{cases} \tag{2.18}$$

and

$$X_{2i} = \begin{cases} 1 & \text{if } t_\ell < T_i \leq t_h \\ 0 & \text{otherwise.} \end{cases} \tag{2.19}$$

The total $(N_1, N_2) = \left(\sum_{i=1}^n X_{1,i}, \sum_{i=1}^n X_{2,i}\right)$ is a trinomial random variable, so $\mathrm{E}[X_1] = p_1 = \Pr(T < t_\ell)$ and $\mathrm{E}[X_2] = p_2 = \Pr(t_\ell \leq T < t_h)$. The unbiased, maximum-likelihood estimator of p_1 is n_1/n, where n_1 is the number of $T_i \leq t_\ell$ and that of p_2 is n_2/n, where n_2 is the number of T_i with $t_\ell < T_i \leq t_h$. Furthermore, $\mathrm{Var}(p_1) = [p_1(1-p_1)]/n$, $\mathrm{Var}(p_2) = [p_2(1-p_2)]/n$, and $\mathrm{Cov}(p_1, p_2) = -p_1 p_2/n$. Because $p_\ell = p_1$ and $p_h = p_1 + p_2$ are continuous functions of p_1 and p_2, the invariance property of maximum-likelihood estimators implies

$$\hat{p}_\ell = \frac{n_1}{n} \tag{2.20}$$

$$\hat{p}_h = \frac{n_1 + n_2}{n} \tag{2.21}$$

$$\widehat{\mathrm{Var}}(\hat{p}_\ell) = \frac{\hat{p}_\ell\left(1 - \hat{p}_\ell\right)}{n} \tag{2.22}$$

$$\widehat{\mathrm{Var}}(\hat{p}_h) = \frac{\hat{p}_h\left(1-\hat{p}_h\right)}{n} \tag{2.23}$$

$$\widehat{\mathrm{Cov}}(\hat{p}_\ell, \hat{p}_h) = \frac{\hat{p}_\ell\left(1-\hat{p}_h\right)}{n}. \tag{2.24}$$

Estimates of bounds on mean timeliness Assume the analyst wishes to screen out systems for which $\mathrm{E}[\mathcal{T}(T)] < \mathrm{E}_{\mathrm{MIN}}$. Again, from (2.7)

$$\begin{aligned}\mathrm{E}[\mathcal{T}(T)] &= -\int_0^{t_\ell} \mathrm{F}(t)d\mathcal{T}(t) - \int_{t_\ell}^{t_h} \mathrm{F}(t)d\mathcal{T}(t) - \int_{t_h}^{\infty} \mathrm{F}(t)d\mathcal{T}(t) \\ &\leq p_\ell\left(1-\mathcal{T}_\ell\right) + p_h\left(\mathcal{T}_\ell - \mathcal{T}_h\right) + \mathcal{T}_h. \end{aligned} \tag{2.25}$$

Because (2.25) is a linear function of (p_ℓ, p_h), it follows that a maximum-likelihood estimator for the upper bound on mean timeliness is

$$\hat{\mathrm{E}}_2^+\left[\mathcal{T}(T)\middle|\mathcal{T}_\ell, \mathcal{T}_h, \hat{p}_\ell, \hat{p}_h\right] = \left(1-\mathcal{T}_\ell\right)\hat{p}_\ell + \left(\mathcal{T}_\ell - \mathcal{T}_h\right)\hat{p}_h + \mathcal{T}_h. \tag{2.26}$$

An approximate upper limit follows from the asymptotic normal distribution of maximum-likelihood estimators [28], with

$$\begin{aligned}\widehat{\mathrm{Var}}\left[\hat{\mathrm{E}}_2^+\right] &= (1-\mathcal{T}_\ell)^2\frac{\hat{p}_\ell(1-\hat{p}_\ell)}{n} + (\mathcal{T}_\ell - \mathcal{T}_h)^2\frac{\hat{p}_h(1-\hat{p}_h)}{n} + \\ &\quad 2(1-\mathcal{T}_\ell)(\mathcal{T}_\ell - \mathcal{T}_h)\frac{\hat{p}_\ell(1-\hat{p}_h)}{n}. \end{aligned} \tag{2.27}$$

Using similar arguments to that for E_2^+ shows that

$$\begin{aligned}\hat{\mathrm{E}}_2^-\left[\mathcal{T}(T)\middle|\mathcal{T}_\ell, \mathcal{T}_h, \hat{p}\ell, \hat{p}_h\right] &= \mathcal{T}_h\hat{p}_h + \left(\mathcal{T}_\ell - \mathcal{T}_h\right)\hat{p}_\ell \\ &= \hat{\mathrm{E}}_1^-\left[\mathcal{T}(T)\middle|\mathcal{T}_\ell, \hat{p}\ell\right] + \hat{p}_2\mathcal{T}_h \end{aligned} \tag{2.28}$$

is a maximum-likelihood estimator of the lower bound on $\mathrm{E}[\mathcal{T}(T)]$ with

$$\widehat{\mathrm{Var}}\left[\mathrm{E}_2^-\right] = \left(\mathcal{T}_\ell - \mathcal{T}_h\right)^2\frac{\hat{p}_\ell(1-\hat{p}_\ell)}{n} + \mathcal{T}_h^2\frac{\hat{p}_h(1-\hat{p}_h)}{n} + 2\mathcal{T}_h\left(\mathcal{T}_\ell - \mathcal{T}_h\right)\frac{\hat{p}_\ell(1-\hat{p}_h)}{n}$$

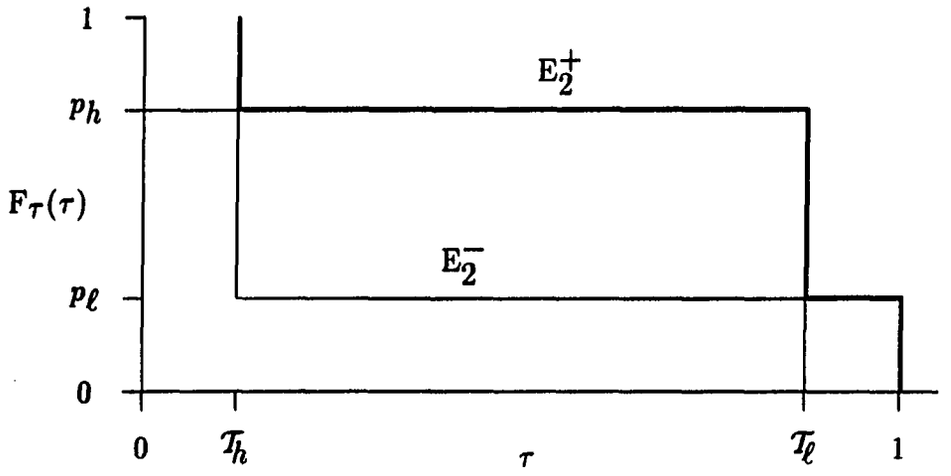


Figure 2.4: Two-Estimate Bounds on E[$T(T)$] for an Individual Needline

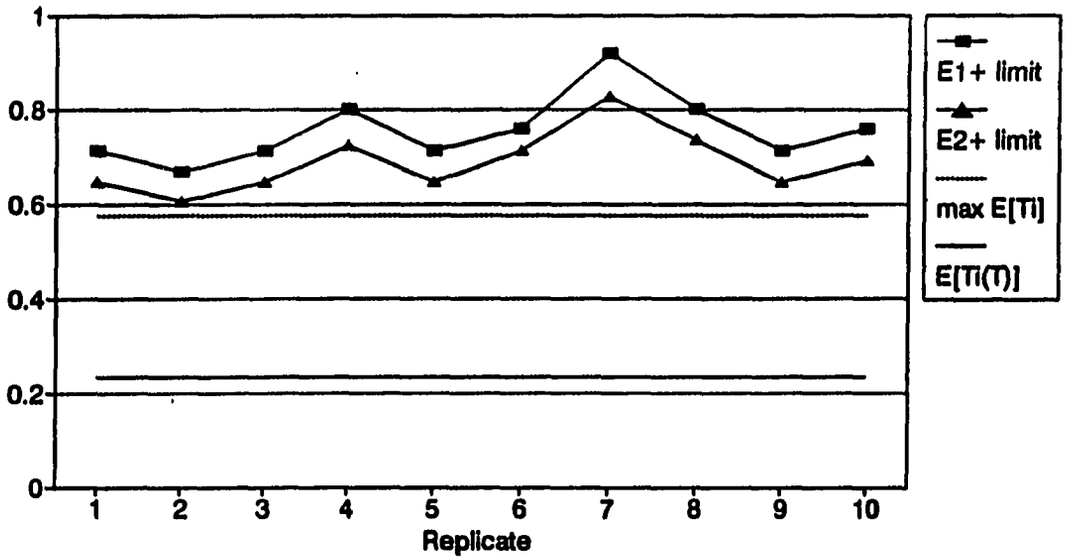
$$= \widehat{\text{Var}}\left[\text{E}_1^-\right] + \mathcal{T}_h^2 \frac{\hat{p}_2(1-\hat{p}_2)}{n} - 2\mathcal{T}_\ell\mathcal{T}_h \frac{\hat{p}_\ell(1-\hat{p}_2)}{n} \tag{2.29}$$

Figure 2.4 illustrates the two-probability bounds on E[$T(T)$]. Figures 2.5 and 2.6 show some sample estimates of E^+ and E^-, respectively, using ten independent replicates of twenty observations each. The solid line in Figure 2.5 is the theoretical upper bound (2.10), while that in Figure 2.6 is the theoretical lower bound (2.11).

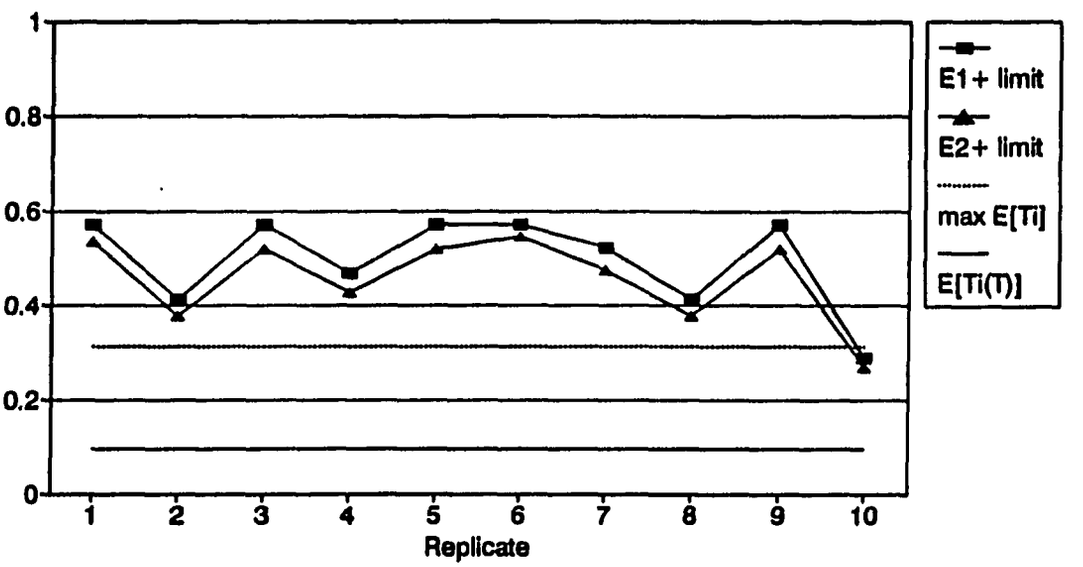
The effect of $\mathcal{T}_\ell$ and $\mathcal{T}_h$ on estimates The primary concerns in the choice of $\mathcal{T}_\ell$ and $\mathcal{T}_h$ are 1) the ability of the C^2 community to establish their values and 2) the error in estimating (2.5). However, it is good to check the effect of these choices on bounds. First of all,

$$\hat{\text{E}}_{1\alpha}^+ - \hat{\text{E}}_{1\alpha}^- = \left(1-\mathcal{T}_h\right)\left(\hat{p}_{h\alpha}^+ - \hat{p}_{\ell\alpha}^-\right) + \left(1-\mathcal{T}_h-\mathcal{T}_\ell\right)\hat{p}_{\ell\alpha}^- + \mathcal{T}_h. \tag{2.30}$$

Choosing $\mathcal{T}_h = 1 - \mathcal{T}_\ell$ causes this difference to be independent of $\hat{p}_{\ell\alpha}^-$ and depend on events within $\Omega = (t_\ell, t_h)$. Choosing $\mathcal{T}_h$ small, reduces the size of the difference and

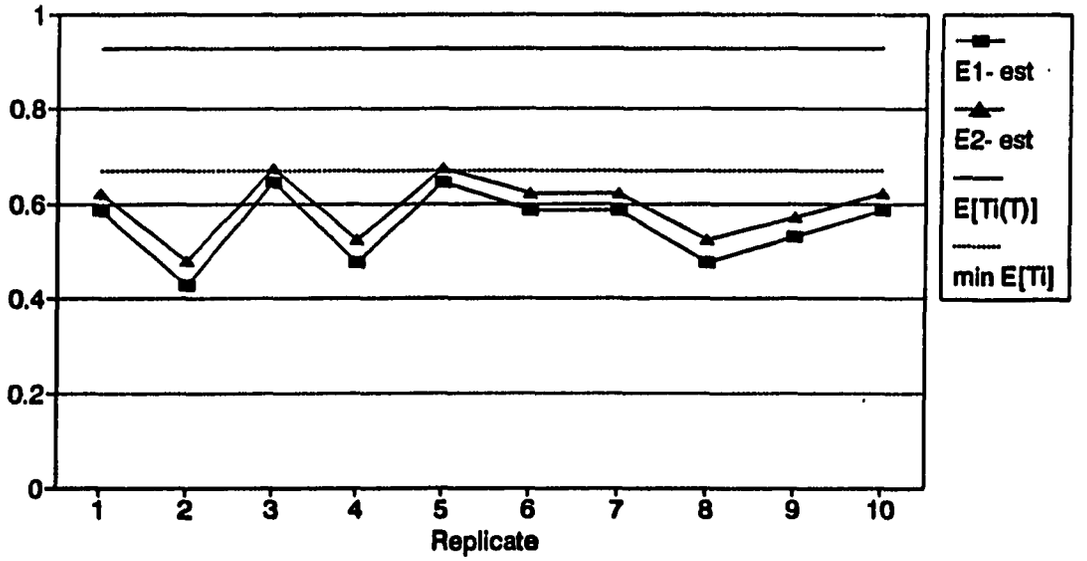


a. With $\lambda = 2.8$, $p_\ell = 0.018$ and $p_h = 0.594$

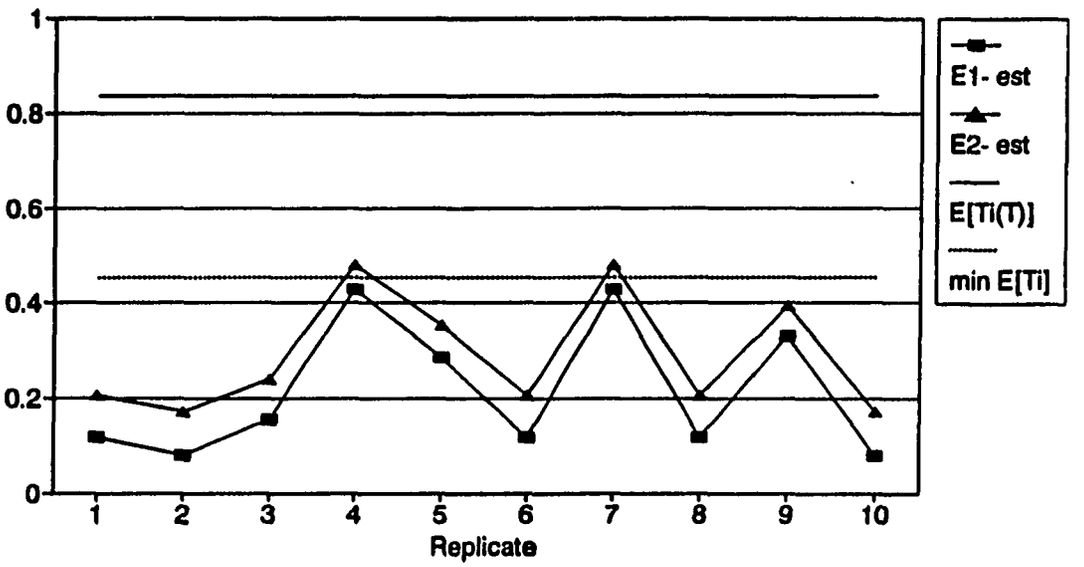


b. With $\lambda = 2.0$, $p_\ell = 0.264$ and $p_h = 1.000$

Figure 2.5: Estimates of the Upper Bound on $\mathrm{E}[\mathcal{T}(T)$ in Ten Replicates of Twenty Independent Observations Each, Knowing Only $\mathcal{T}_\ell = 0.9$ and $\mathcal{T}_h = 0.1$



a. With $\lambda = 0.5$, $p_\ell = 0.713$ and $p_h = 1.000$



b. With $\lambda = 1.5$, $p_\ell = 0.442$ and $p_h = 1.000$

Figure 2.6: Estimates of the Lower Bound on $\mathrm{E}[\mathcal{T}(T)$ in Ten Replicates of Twenty Independent Observations Each, Knowing Only $\mathcal{T}_\ell = 0.9$ and $\mathcal{T}_h = 0.1$

further emphasizes the events within Ω. Assuming $\mathcal{T}_h = 1 - \mathcal{T}_\ell$,

$$\begin{aligned}\hat{\mathrm{E}}^+_{2\alpha} - \hat{\mathrm{E}}^-_{2\alpha} &= \left(1-\mathcal{T}_h\right)\left(\hat{p}^+_{h\alpha} - \hat{p}^-_{\ell\alpha}\right) - \\ &\quad \left(1-\mathcal{T}_\ell\right)\left(\hat{p}^+_{h\alpha} - \hat{p}^+_{\ell\alpha}\right) - \mathcal{T}_h\left(\hat{p}^-_{h\alpha} - \hat{p}^-_{\ell\alpha}\right) + \mathcal{T}_h. \qquad (2.31)\end{aligned}$$

Noting that $\hat{p}^+_{h\alpha} - \hat{p}^+_{\ell\alpha} \approx \hat{p}^-_{h\alpha} - \hat{p}^-_{\ell\alpha}$,

$$\hat{\mathrm{E}}^+_{2\alpha} - \hat{\mathrm{E}}^-_{2\alpha} \approx \left(1-\mathcal{T}_h\right)\left(\hat{p}^+_{h\alpha} - \hat{p}^-_{\ell\alpha}\right) + \mathcal{T}_h\left[1 - 2\left(\hat{p}^+_{h\alpha} - \hat{p}^+_{\ell\alpha}\right)\right]. \qquad (2.32)$$

This means that unless $p_h - p_\ell > 0.5$, a large $\mathcal{T}_h$ will cause the width of the interval of uncertainty to actually increase. Chapter 3 argues that $p_h - p_\ell$ is usually smaller than 0.5. Thus, in selecting $\mathcal{T}_\ell$ and $\mathcal{T}_h$, one should 1) select $\mathcal{T}_h = 1 - \mathcal{T}_\ell$ and 2) make $\mathcal{T}_h$ as small as practical, subject of course to the primary concerns above. The assignment $\mathcal{T}_\ell = 0.9$ and $\mathcal{T}_h = 0.1$ is consistent with these recommendations.

The hybrid estimator of timeliness

The estimators above make no assumptions about the nature of either $\mathrm{F}(t)$ or $\mathcal{T}(t)$, except at t_ℓ and t_h. The final estimator assumes a particular distribution for T. Comparing Figure 2.7 to Figure 1.3 on Page 17, implies that, unless φ is very small, $\mathrm{E_H}[\mathcal{T}(T)]$ is a good estimator of $\mathrm{E}[\mathcal{T}(T)]$. As shown in Chapter 3, this distribution is a robust characterization of $\mathrm{F}(t)$ for estimating $\mathrm{E}[\mathcal{T}(T)]$ over a wide range of actual distributions. Of course, this approximation is appropriate only when $\mathcal{T}(t)$ is known.

The hybrid distribution As depicted in Figure 2.8, $\mathrm{F_H}(t)$ assumes a uniform distribution on $[0, t_\ell)$ and an independent exponential distribution with mean

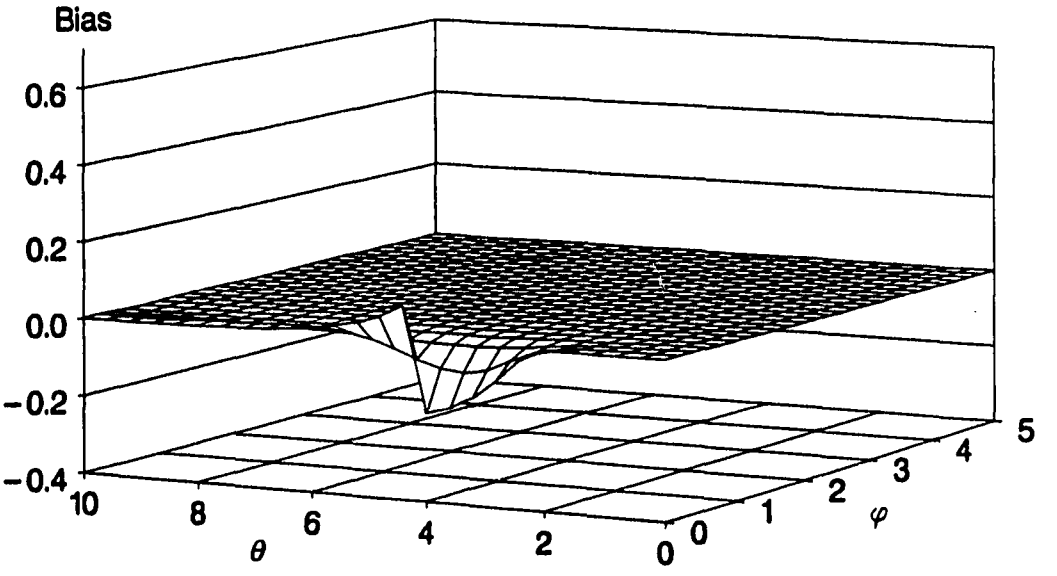


Figure 2.7: Error Estimating $E[\mathcal{T}_2(T;1.0,2.0)]$ by $E_H[\mathcal{T}(T)]$ when T Follows a Gamma Distribution having Mean θ and Coefficient of Variation φ

ζ for times greater than t_ℓ on $[t_\ell, \infty]$.

$$F_H\left(t; t_\ell, p_\ell, \zeta\right) \triangleq \begin{cases} \left(\frac{p_\ell}{t_\ell}\right) t & 0 \le t < t_\ell \\ 1 - \left(1 - p_\ell\right) \exp\left[-\left(\frac{t - t_\ell}{\zeta}\right)\right] & t \ge t_\ell. \end{cases} \tag{2.33}$$

Its density, shown in Figure 2.9, is:

$$f_H\left(t; t_\ell, p_\ell, \zeta\right) \triangleq \begin{cases} \frac{p_\ell}{t_\ell} & 0 \le t < t_\ell \\ \left(1 - p_\ell\right)\left(\frac{1}{\zeta}\right) \exp\left[-\left(\frac{t - t_\ell}{\zeta}\right)\right] & t \ge t_\ell. \end{cases} \tag{2.34}$$

Note that p_ℓ determines $f_H(t)$ on $[0, t_\ell)$ and the conditional probability

$$\Pr\left\{T \le t \middle| T > t_\ell\right\} = 1 - \exp\left[-\left(\frac{t - t_\ell}{\zeta}\right)\right] \tag{2.35}$$

depends only on ζ.

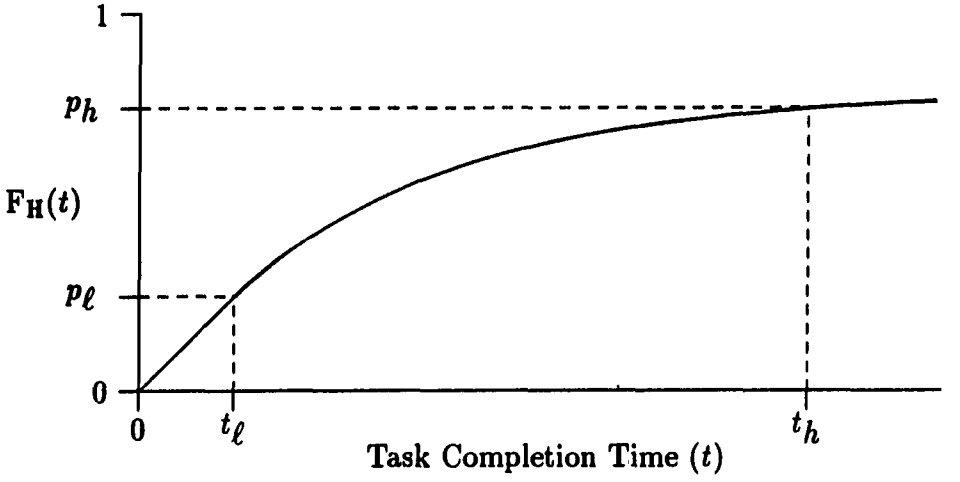


Figure 2.8: The Hybrid Cumulative-Distribution Function

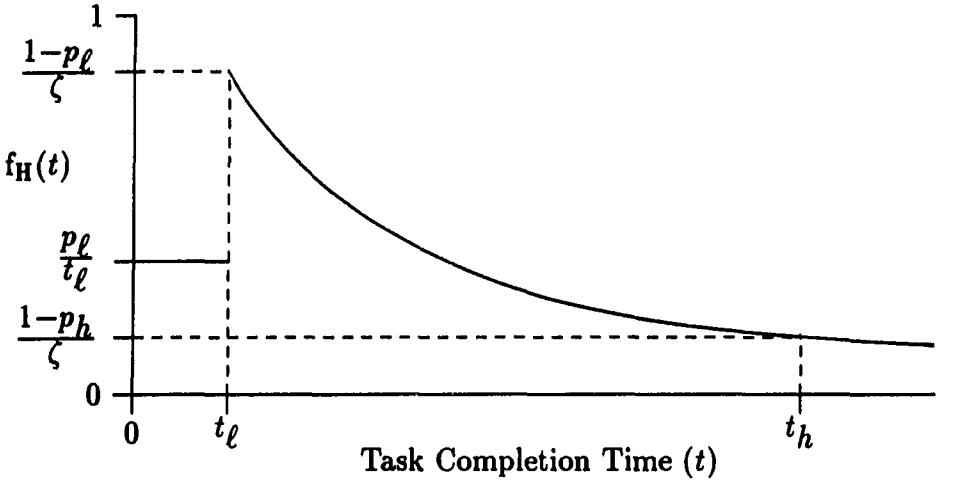


Figure 2.9: The Hybrid Density Function

Estimators of the hybrid parameters Assume $\{t_i; i = 1,\ldots,n\}$ is a random sample from the hybrid time distribution. Also, without loss of generality, assume that the sample as been re-ordered so the first n_1 of the t_i are in $[0, t_\ell)$, the next n_2 in $[t_\ell, t_h)$ and the last $n_3 = n - n_1 - n_2$ are in $[t_h, \infty)$. To focus on times in (t_ℓ, t_h), the estimator assumes times greater than t_h are censored[1] at time t_h. Additionally, since the timeliness of cancelled calls is zero, such calls are counted as if they were completed in $[t_h, \infty]$. The likelihood of the sample is [28]:

$$\begin{aligned} L\left(p_\ell, \zeta \left| \{t_i\}, t_\ell, t_h \right.\right) &= \prod_{i=1}^{n_1+n_2} \mathrm{f_H}\left(t_i; p_\ell, \zeta\right) \times \left[1 - \mathrm{F_H}(t_h; p_\ell, \zeta)\right]^{n_3} \\ &= \left(\frac{p_\ell}{t_\ell}\right)^{n_1} \prod_{i=n_1+1}^{n_1+n_2} \left(\frac{1-p_\ell}{\zeta}\right) \exp\left[-\left(\frac{t_i - t_\ell}{\zeta}\right)\right] \times \\ &\quad \left\{\left(1-p_\ell\right) \exp\left[-\left(\frac{t_h - t_\ell}{\zeta}\right)\right]\right\}^{n_3}. \end{aligned} \tag{2.36}$$

The partial derivatives of the log-likelihood function are:

$$\frac{\partial \ln L(\cdot)}{\partial p_\ell} = \frac{n_1}{p_\ell} - \frac{n - n_1}{1 - p_\ell} \tag{2.37}$$

$$\frac{\partial \ln L(\cdot)}{\partial \zeta} = -\frac{n_2}{\zeta} + \sum_{i=n_1+1}^{n_1+n_2} \frac{t_i - t_\ell}{\zeta^2} + \frac{n_3(t_h - t_\ell)}{\zeta^2} \tag{2.38}$$

Setting these equal to zero and solving them yields:

$$\hat{p}_{\ell_{\mathrm{MLE}}} = \frac{n_1}{n} \tag{2.39}$$

$$\hat{\zeta}_{\mathrm{MLE}} = \frac{T_{\mathrm{OBS}}}{n_2} \tag{2.40}$$

[1] That is, the fact that $t_i > t_h$ is used, but not the value of t_i itself.

where

$$T_{\text{OBS}} \triangleq \sum_{i=n_1+1}^{n_1+n_2} \left(t_i - t_\ell\right) + n_3 \left(t_h - t_\ell\right). \tag{2.41}$$

is the total of the "observed[2]" times greater than t_ℓ. Because $\partial^2 \ln L(\cdot)/\partial p_\ell^2$ and $\partial^2 \ln L(\cdot)/\partial \zeta^2$ are non-positive, (2.39) and (2.40) are maximum-likelihood estimators.

Clearly, $\hat{p}_{\ell_{\text{MLE}}}$ is unbiased. Also,

$$\begin{aligned} \mathrm{E}\left[\hat{\zeta}_{\text{MLE}}\right] &= \mathrm{E}\left[\frac{T_{\text{OBS}}}{n_2}\right] \\ &= \mathrm{E}\left[\sum_{i=n_1+1}^{n_1+n_2} \frac{t_i - t_\ell}{n_2}\right] + \mathrm{E}\left[\left(\frac{n_3}{n_2}\right)\left(t_h - t_\ell\right)\right]. \end{aligned} \tag{2.42}$$

The first term of (2.42) is the mean value of $T - t_\ell$ on $[t_\ell, t_h)$, so

$$\begin{aligned} \mathrm{E}\left[\sum_{i=n_1+1}^{n_1+n_2} \frac{t_i - t_\ell}{n_2}\right] &= \frac{\int_{t_\ell}^{t_h} \left(x - t_\ell\right)\left(1 - p_\ell\right)\left(\frac{1}{\zeta}\right) \exp\left(-\frac{x - t_\ell}{\zeta}\right) dx}{p_h - p_\ell} \\ &= \frac{-\left(t_h - t_\ell\right)\left(1 - p_\ell\right) \exp\left(-\frac{t_h - t_\ell}{\zeta}\right)}{p_h - p_\ell} + \\ &\quad \frac{\zeta\left(1 - p_\ell\right)\left[1 - \exp\left(-\frac{t_h - t_\ell}{\zeta}\right)\right]}{p_h - p_\ell}. \\ &= -\left(t_h - t_\ell\right)\left(\frac{1 - p_h}{p_h - p_\ell}\right) + \zeta\left(\frac{p_h - p_\ell}{p_h - p_\ell}\right). \end{aligned} \tag{2.43}$$

Because the second term of (2.42) is $\left(t_h - t_\ell\right)\left(1 - p_h\right) / \left(p_h - p_\ell\right)$, $\hat{\zeta}_{\text{MLE}}$ is also unbiased[3]. Since $\hat{p}_{\ell_{\text{MLE}}}$ is the same as those used earlier and $\hat{\zeta}_{\text{MLE}}$ is the only

[2]Cancelled calls are treated as they completed at time t_h.

[3]See p. 80 for an alternate argument.

estimator of ζ that will be used in this paper, from now on they will be referred to as simply $\hat{p}_\ell$ and $\hat{\zeta}$, respectively.

Again, one can use the asymptotically normal property of $\hat{p}_\ell$ to find intervals for p_ℓ. Although $\hat{\zeta}$ is also asymptotically normal, it converges very slowly. [28, §3.2] discusses numerous approximate distributions for estimates of the exponential distribution under this type of censoring. A good, but simple one is to treat $2n_2\hat{\zeta}/\zeta$ as approximately $\chi^2_{(2n_2+1)}$. This gives good coverage with samples as small as ten. Thus an approximate $(1-\alpha/2)\times 100\%$ lower c.l. for ζ would be $2n_2\hat{\zeta}/\chi^2_{(2n_2+1),\alpha/2}$.

An estimator of mean timeliness Since $\mathcal{T}(t)$ is assumed known, the hybrid estimate of timeliness is simply the expectation,

$$\begin{aligned}
&\mathrm{E_H}[\mathcal{T}(T)|p_\ell,\zeta,t_\ell] \\
&= \int_0^\infty \mathcal{T}(t)\, df_{\mathrm{H}}(t;p_\ell,\zeta,t_\ell)\, dt \\
&= p_\ell \int_0^{t_\ell} \frac{\mathcal{T}(t)}{t_\ell}\, dt \;+\; \left(1-p_\ell\right) \int_{t_\ell}^\infty \frac{\mathcal{T}(t)}{\zeta} \exp\left[-\left(\frac{t-t_\ell}{\zeta}\right)\right] dt. \\
&= p_\ell m\left[\mathcal{T}(T)\,\middle|\,T<t_\ell\right] \;+\; \left(1-p_\ell\right) \mathrm{E_M}\left[\mathcal{T}(T+t_\ell)\middle|\,\zeta\right] \qquad (2.44)
\end{aligned}$$

where $m\left[\mathcal{T}(T)\,\middle|\,T<t_\ell\right]$ is the mean value of $\mathcal{T}(t)$ on $[0,t_\ell)$ and $\mathrm{E_M}\left[\mathcal{T}(T+t_\ell)\middle|\,\zeta\right]$ is the mean timeliness, assuming T is an exponential random variable with mean ζ and the timeliness function is $\mathcal{T}(t+t_\ell)$. Because $\mathcal{T}(t)$ is a non-increasing function, it follows that (2.44) increases with increasing p_ℓ. Increasing ζ causes $\mathrm{E_M}[\cdot]$ decrease. Hence $\mathrm{E_H}\left[\mathcal{T}(T)\,\middle|\,p_\ell,\zeta\right]$ is an increasing function of p_ℓ and a decreasing function of ζ.

Being a function of maximum-likelihood functions, $\hat{\mathrm{E}}[\mathcal{T}(T)]$ is $\mathrm{E}\left[\mathcal{T}(T)\,\middle|\,\hat{p}_\ell,\hat{\zeta},t_\ell\right]$. An approximate upper $(1-\alpha)\times 100\%$ for $\mathrm{E_H}[\mathcal{T}(T)]$ can be computed as follows.

1. Noting that $\mathrm{E_H}[\mathcal{T}(T)]$ is an increasing function of p_ℓ and that the marginal likelihood of p_ℓ is not affected by the choice of ζ, find $p^+_{\ell\,\alpha/2}$, a $(1-\alpha/2)\times 100\%$ upper c.l. for p_ℓ using the entire sample.

2. Noting that $\mathrm{E_H}[\mathcal{T}(T)]$ is a decreasing function of ζ and that, given p_ℓ, the conditional distribution of T on $[t_\ell, \infty)$ is not affected by p_ℓ, find $\zeta^-_{\alpha/2}$, a $(1-\alpha/2)\times 100\%$ lower c.l. for ζ, using only the observed $t_i > t_\ell$ and cancelled calls.

3. Treat $\mathrm{E_H}\left[\mathcal{T}(T)\,\middle|\,p^+_{\ell\,\alpha/2}, \zeta^-_{\alpha/2}, t_\ell\right]$ as the upper $(1-\alpha)\times 100\%$ c.l. for $\mathrm{E_H}[\mathcal{T}(T)]$.

One can use similar procedures to approximate lower and two-sided c.i.s for $\mathrm{E_H}[\mathcal{T}(T)]$.

An example Substituting the $\mathcal{T}_2(t;\tilde{t},t_0)$ timeliness model into (2.44) yields:

$$\mathrm{E_H}[\mathcal{T}(T;\tilde{t},t_0)|p_\ell,\zeta,\mathcal{T}_\ell] \;=\; p_\ell\, C_1\left(\tilde{t},t_0,\mathcal{T}_\ell\right) \;+\; \left(1-p_\ell\right)\mathcal{T}_\ell\left(\frac{\tilde{t}}{\zeta+\tilde{t}}\right) \qquad (2.45)$$

where

$$C_1\left(\tilde{t},t_0,\mathcal{T}_\ell\right) \;=\; \frac{t_0+\tilde{t}(1-\mathcal{T}_\ell)}{t_0-\tilde{t}\ln\mathcal{T}_\ell} \qquad (2.46)$$

is the mean value of $\mathcal{T}_2(t)$ on $[0,t_\ell)$. One may also use (A.2) and (A.3) to express the expectation in terms of t_ℓ, t_h, $\mathcal{T}_\ell$, V_h, p_ℓ and ζ, or equivalently, $\mathcal{P}$, $\mathcal{T}_\ell$, $\mathcal{T}_h$, p_ℓ and ζ. Table 2.1 lists values of $C_1(\mathcal{P};\mathcal{T}_\ell,\mathcal{T}_h)$ for all twelve $\mathcal{P}$ codes with several values of $\mathcal{T}_\ell$ and $\mathcal{T}_h$.

Figure 2.10 illustrates approximate 80% confidence intervals on $\mathrm{E}[\mathcal{T}(T)]$ in the two-link path of Example 1.3. The intervals appear to be conservative, which is not surprising, considering the approximations used. They may not be as generous as they seem, however, since the distribution in Example 1.3 is not very unusual, and

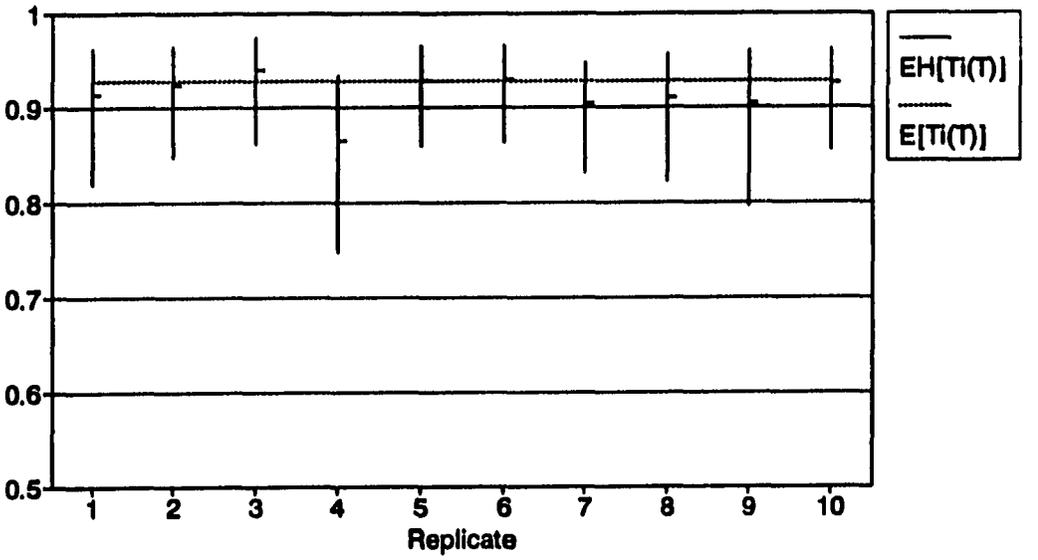
Table 2.1: Coefficient $C_1(\cdot)$ for the CDB Perishability Codes

	$C_1(\mathcal{P}; \mathcal{T}_\ell, \mathcal{T}_h)$		
$\mathcal{P}$	$\mathcal{T}_\ell = 0.90$ $\mathcal{T}_h = 0.10$	$\mathcal{T}_\ell = 0.90$ $\mathcal{T}_h = 0.50$	$\mathcal{T}_\ell = 0.95$ $\mathcal{T}_h = 0.10$
0	0.9951	0.9818	0.9989
1	0.9976	0.9909	0.9994
2	0.9992	0.9970	0.9998
3	0.9988	0.9954	0.9997
4	0.9976	0.9909	0.9994
5	0.9878	0.9544	0.9971
6	0.9780	0.9179	0.9948
7	0.9966	0.9872	0.9992
8	0.9969	0.9884	0.9993
9	0.9971	0.9891	0.9993
A	0.9902	0.9635	0.9977
B	0.9491	0.9491	0.9748

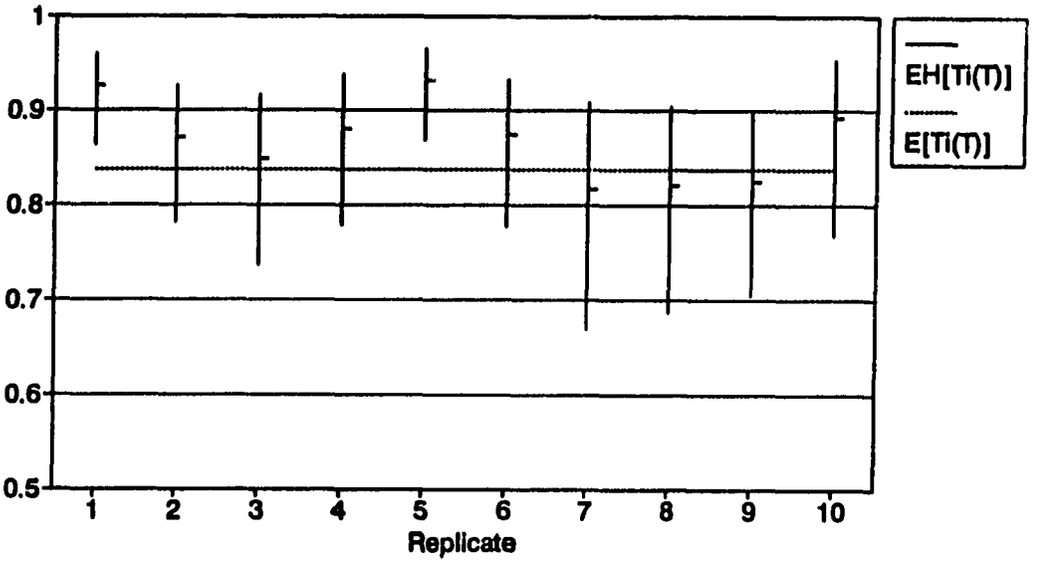
these intervals are designed for many cases. Figure 2.11 shows approximate 80% intervals for ten replicates of a simulation in a (M/D/1) system. The hybrid estimate has two intervals excluding the true mean and one nearly so, which is not surprising. Note that the nice, tight intervals based on E[T] alone in Part b are definitely biased.

Some general comments on estimators

Common properties All proposed estimators are robust and unbiased. In addition, being maximum-likelihood estimators, they possess the useful invariance property [29]. An approximate conservative confidence interval is shown for each, more exact intervals are suggested below. The primary objective in each case is to represent F(t) in such a way that E[$\mathcal{T}(T)$] point estimates are unbiased and interval estimates are reasonably accurate.

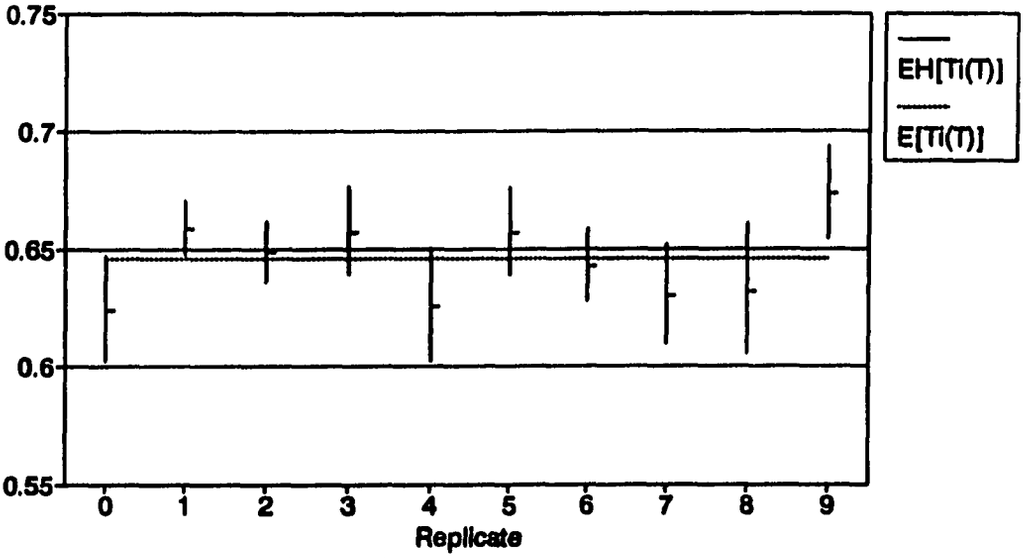


a. With $\lambda = 0.5$ and $\mathcal{P} = 6$

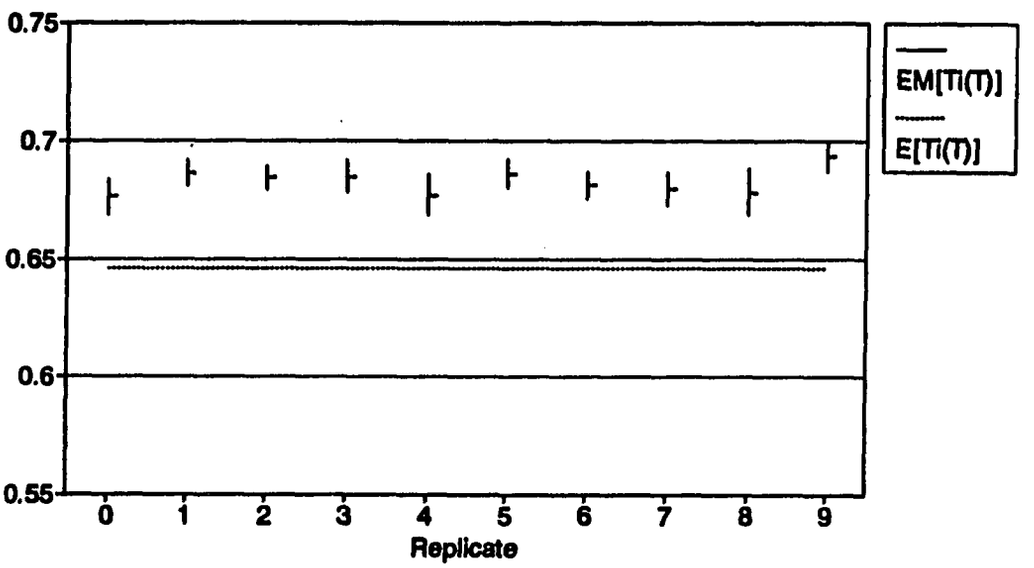


b. With $\lambda = 1.5$ and $\mathcal{P} = 6$

Figure 2.10: Approximate 80% Confidence Intervals on $\mathrm{E}[\mathcal{T}(T)]$ in Ten Replicates of Twenty Independent Observations Each from the Two-Link Path in Example 1.3, Using $\mathrm{E}_{\mathrm{II}}[\mathcal{T}(T)]$



a. Using the Hybrid Estimator



b. Using the Markovian Estimator

Figure 2.11: Approximate 80% Confidence Intervals on $E[\mathcal{T}(T)]$ in Ten Replicates of Twenty to Forty Observations Each of a Simulation of a (M/D/1) Queue Using the Indicated Estimator

Robustness It is important that the method is robust. Probability estimates are completely insensitive to the shape of F(t), except at the selected point(s). In particular, if $T = \sum j = 1^m \lambda_j t_j$ is a mixture of several distributions, the probability estimates for T will still be accurate. Chapter 3 shows $(\hat{p}_\ell, \hat{\zeta})$ has similar properties when used with the hybrid estimate.

Because of this robustness, analysts need not worry about distinct distributions that arise from different event sequences within a given task, such as completion on the first call, completion on the second call, etc. Additionally, from (2.2),

$$\mathrm{E}_S[\mathcal{V}(T)] \;=\; \sum_{k=1}^{c} I_k \left(\frac{\lambda_k}{\lambda}\right) \mathrm{E}\left[\mathcal{T}_k\left(t_k\right)\right] \tag{2.47}$$

where $k = 1,\ldots,c$ represents equivalence classes determined by importance and timeliness function values. Here I_k and $\mathcal{T}_k(t)$ are the common importance weight and timeliness function, respectively, for equivalence class k. Robustness in the face of such complex time distributions for the T_k can lead to a great savings in run-time data collection. In the CDB, for example, while there are hundreds of needlines, there are only thirty-six distinct COF and $\mathcal{P}$ combinations.

Confidence intervals The c.i.s shown are approximate and should be used carefully. In determining estimator c.i.s, one should bear in mind that the end goal is to find good c.i.s for E[$\mathcal{T}(T)$]. For example, the upper $(1-\alpha) \times 100\%$ confidence limit for E[$\mathcal{T}(T)$ is $\mathrm{E}\left[\mathcal{T}(T)\,\middle|\,\Theta^\star\right]$ where $\Theta^\star$ is the solution to

$$\begin{aligned} \max \quad & \mathrm{E}\left[\mathcal{T}(T)\,\middle|\,\tilde{\Theta}\right] \\ \text{subject to} \quad & \Pr\left\{\Theta \in \mathcal{S}\left(\tilde{\Theta}\right)\right\} = 1-\alpha \\ & \tilde{\Theta} \in \mathcal{S} \end{aligned} \tag{2.48}$$

where $\tilde{\Theta}$ is a particular set of parameter values, $\mathcal{S}\left(\tilde{\Theta}\right)$ is a subset of the parameter space $\mathcal{S}$ bounded by $\tilde{\Theta}$. For example, a $(1-\alpha)\times 100\%$ upper confidence limit for E_1^+, is $\mathrm{E}_1\left[\mathcal{T}(T)\middle|p_h^\star\right]$, where $p_h^\star$ is the solution to

$$\begin{aligned} \max \quad & \mathrm{E}\left[\mathcal{T}(T)\middle|\mathcal{T}_h,\tilde{p}_h\right] \\ \text{subject to} \quad & \Pr\left\{p_h \le \tilde{p}_h\right\} = 1-\alpha \\ & 0 \le \tilde{p} \le 1. \end{aligned} \tag{2.49}$$

In this case the problem is more simply stated as

$$\begin{aligned} \max \quad & \tilde{p}_h \\ \text{subject to} \quad & \sum_{i=0}^{n_1+n+2} \binom{n}{i} \tilde{p}_h^i \left(1-\tilde{p}_h\right)^{n_1+n_2-i} = 1-\alpha \\ & 0 \le \tilde{p} \le 1. \end{aligned} \tag{2.50}$$

One can find an exact interval, searching the curve

$$\sum_{i=0}^{n_1+n+2} \binom{n}{i} \tilde{p}_h^i \left(1-\tilde{p}_h\right)^{n_1+n_2-i} = 1-\alpha \tag{2.51}$$

for the point at which $\tilde{p}_h$ is maximum.

When two parameters are involved, the probability constraints are more complex and one must conduct a two-dimensional search for the optimum. For example, an exact upper limit for E_2^+ would be $\mathrm{E}_2^+\left[\mathcal{T}(T)\middle|\mathcal{T}_\ell,\mathcal{T}_h,p_\ell^\star,p_h^\star\right]$, where $(p_\ell^\star,p_h^\star)$ is the solution to

$$\max \quad (1-\mathcal{T}_\ell)\tilde{p}_\ell + (\mathcal{T}_\ell - \mathcal{T}_h)\tilde{p}_h$$

$$\text{subject to:} \quad \Pr\left\{p_\ell \le \tilde{p}_\ell, p_h \le \tilde{p}_h\right\} = 1-\alpha$$

$$0 \le p_\ell \le 1$$

$$\tilde{p}_\ell \le \tilde{p}_h \le 1. \tag{2.52}$$

This entails a numerical search of the curve

$$\sum_{i=0}^{n_1}\sum_{j=0}^{n_2}\binom{n}{i,j,n-i-j}\tilde{p}_\ell^{\,i}\left(\tilde{p}_h-\tilde{p}_\ell\right)^j\left(1-\tilde{p}_h\right)^{n-i-j} \;=\; 1-\alpha$$

for a maximum of the objective function.

Because all estimators are maximum-likelihood estimators, one may replace the probability statement in any of the problems by the likelihood ratio. In general, $-2\ln\left[L\left(\tilde{\theta}\right)/L\left(\hat{\theta}\right)\right] \sim \chi^2_p$, where p is the number of parameters to estimate. Thus, the probability constraint for the E_2^+ problem can be stated

$$n_1\ln\left(\tilde{p}_\ell\right)+n_2\ln\left(\tilde{p}_h-\tilde{p}_\ell\right)+n_3\ln\left(1-\tilde{p}_h\right) \;=\; \mathcal{K}\left(n,n_1,n_2,\alpha\right) \tag{2.53}$$

where

$$\mathcal{K}\left(n,n_1,n_2,\alpha\right) \;=\; n_1\ln\left(n_1\right)+n_2\ln\left(n_2\right)+\left(n-n_1-n_2\right)\ln\left(n-n_1-n_2\right)-$$

$$n\ln(n)-\frac{\chi^2_{2,1-\alpha}}{2} \tag{2.54}$$

is a constant. Although this still entails a two-dimensional search, the function is much simpler. [30] describes methods to solve optimization problems of this type. [28] presents methods of solving likelihood-ratio problems.

Summary

Focused measures of performance provide information that is more relevant to $\mathrm{E}[\mathcal{T}(T)]$ than unfocused measures of performance. The one- and two- probability

bounds provide useful information on $E[\mathcal{T}(T)]$ even though only percentiles of the timeliness function are known. The hybrid estimator will provide more precise estimates of $E[\mathcal{T}(T)]$ when $\mathcal{T}(t)$ is well-defined. Unfortunately, $\mathcal{T}(t)$ and other C^2 specifications are often not precisely known. The next section explores problems caused by such imprecision and presents methods to cope with them.

Pseudo-Confidence Intervals

Acceptable C^3I systems provide adequate service over all reasonable combinations of offered load operating conditions, and models of $\mathcal{V}(t)$. All modeling data is approximate to some degree. A statement that calls occur 20 times a day could mean, for example, "between 10 and 30 with subjective probability 0.95" or "between 15 and 25 with subjective probability 0.99." Even if, as in the CDB, the C^2 parameter variance is not noted, it still exists. That is, as C^2 experts validate modeling data, each has his or her own interval and subjective probability standard. These experts use these unspoken standards in judging the validity of arrival rates, message lengths, etc. The problem is not that these standards exist, but rather that each expert uses different ones. The same can be said about the value functions. For the most part, the emphasis in analysis is on confidence intervals that arise from variation within simulation runs. As a result, although analysts tend to get precise estimates of the model performance, they have results of unknown validity, since the precision of modeling data and effectiveness functions are unknown.

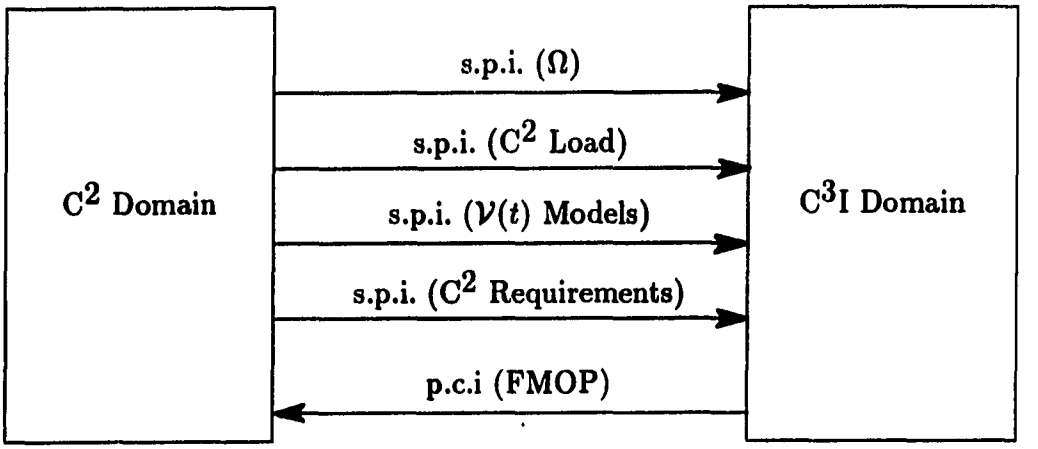


Figure 2.12: How the p.c.i. Enhances Information Flow

The concept

The process starts by assigning intervals of subjective probability to all C^2 parameters and effectiveness functions. This is, of course, a gargantuan task, but as pointed out later, can be guided by C^3I information needs. Large as this task is, it is essential, if one wishes to gauge the validity of a given C^3I analysis. As depicted in Figure 2.12, the C^3I analyst will then report a *pseudo-confidence interval* (p.c.i) for each FMOP to reflect the combined effect of both subjective and statistically-determined probabilities on the results. Like a c.i., the p.c.i is an interval containing the true value with a given probability. However, unlike a true c.i., the p.c.i. is not based on random error alone. For example, suppose a half-width for $E[\mathcal{T}(T)]$ consists of:

$$h = L_\Delta + O_\Delta + S_\Delta + E_\Delta + \mathcal{T}_\Delta + Y_\Delta \tag{2.55}$$

where

L_Δ is the result of uncertainty in the limits on offered load,

O_Δ is the result of uncertainty in the limits on the operational environment,

S_Δ is the result of uncertainty in the C^3I system's characteristics,

E_Δ is the result of uncertainty about the nature of the effectiveness function,

T_Δ is the result of uncertainty about the nature of the timeliness function, and

Y_Δ is the result of random error.

In this case, the first five components of h are due to intervals with subjective probability. Only the last, Y_Δ, is affected by run length. Through factorial design and sensitivity analysis, analysts can determine reasonable estimates of each term of (2.55). Such an analysis will keep their efforts to reduce uncertainty on $E[T(T)]$ in perspective and identify those parts of the over-all problem that needs attention. For example, if $h = 0.25$, $E_\Delta = 0.20$ and $Y_\Delta = 0.02$, one needs to re-examine the effectiveness functions before attempting more simulation runs.

Reflecting the range of the load

Genichi Taguchi is a proponent of exercising models or processes over their expected range of conditions. He is well known for his practical experiment designs and emphasis on total cost. He describes the advantages of his methods in [31] and details his techniques in [32, 33]. Although some of his methods are controversial, his basic premises are not [34, 35, 36, 37].

The basic plan is to estimate performance at selected loads and specific conditions that represent the expected range. From these results, analysts estimate the

response of system performance to changes in the C^2 load and environmental conditions. For example, they may estimate $E[T(T)|b]$, where b is the level of battlefield intensity. However, the primary goal is to represent the full range of C^2 parameter variation. Then, analysts can test C^3I designs and system parameter settings that yield good performance over the probable range of conditions and timeliness functions. To this end, the analyst will use a factorial experiment design [38, 27, 31].

As shown in Figure 2.13, there are two parts to the factorial design. The inner, or parameter, array represents settings of C^3I system parameters the analysts wish to test. The outer, or noise, array represents variations of the offered load and environment. Factors in this array represent extremes of arrival rate, service rate, the arrival and service distributions, environmental interference and projected human behavior. Because it is common practice to replicate the noise matrix at each point of the parameter array, analysts will want to use as few points as possible to represent extremes in the offered load.

It is important to remember that the parameter and noise arrays play opposing roles. The analyst has complete control over the parameter array, since it represents system parameter settings. If he or she does not wish to explore a dimension of control, there is no analytical problem. On the other hand, the analyst has little control over the dimensions and levels of the noise array. For the most part, they represent factors that are either uncontrollable or uneconomical to control. Here the analyst can attempt to simplify the situation as much as possible, but cannot leave out a dimension simply because it complicates the analysis.

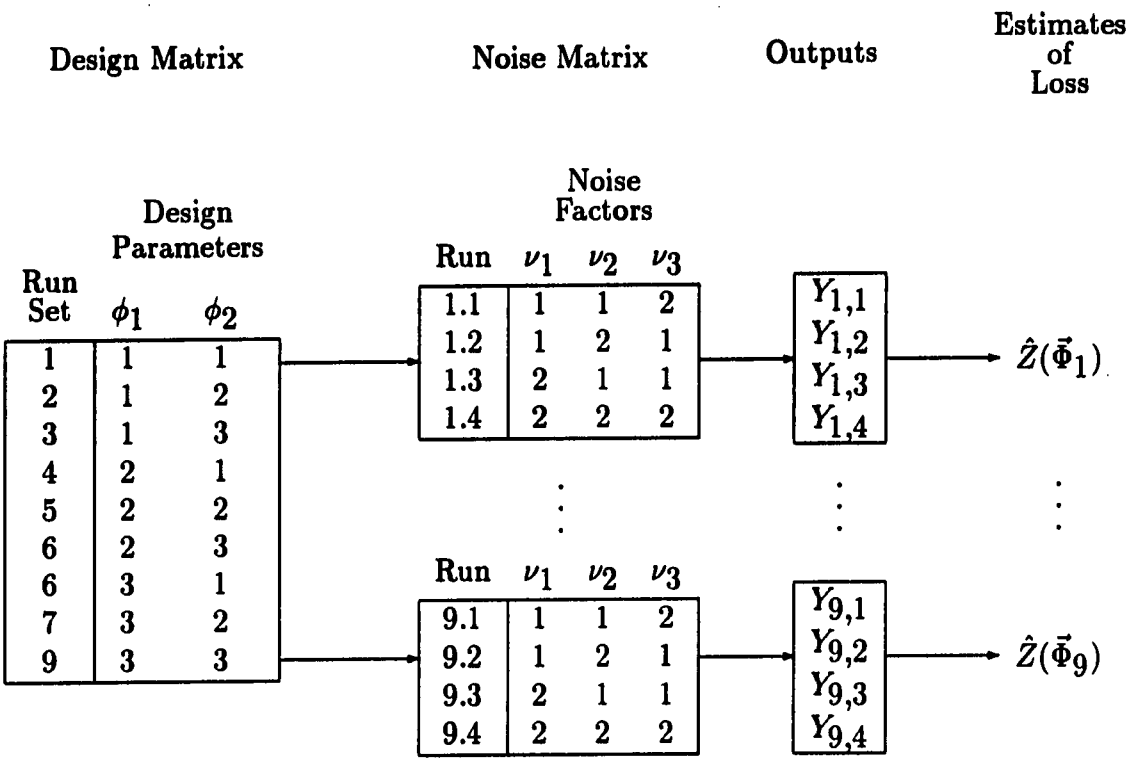


Figure 2.13: An Example of a Taguchi Experiment Design

Reflecting effectiveness function variation

Taguchi's method assumes the effectiveness function is well-defined. In the case of C^3I analysis, however, it is not. Under the timeliness formulation above, there are three sources of variation in effectiveness. The first is variation in the I_i values. The second is variation in the λ_i. Both of these affect the weighting in (2.3). The third is variation in $\mathcal{T}(t)$, the timeliness function. The effect of the first two can be evaluated by performing a sensitivity analysis on (2.3). The last, however, must be handled at simulation time, since it may affect the estimates of the individual $\mathrm{E}[\mathcal{T}_j(T_i)]$.

In the screening phase, the only significant effect of changes in $\mathcal{T}(t)$ would be those that effect t_ℓ or t_h. Thus, there are two classes of variation in $\mathcal{T}(t)$. One possible way to structure this is to first decide on intervals for t_ℓ and t_h, then independently, find intervals for shape parameters of $\mathcal{T}(t)$. Another is to decide on possible models of $\mathcal{T}(t)$, then define values of $t_\ell = \mathcal{T}^{-1}(\mathcal{T}_\ell)$ and $t_h = \mathcal{T}^{-1}(\mathcal{T}_h)$.

Since the bounds on $\mathrm{E}[\mathcal{T}(T)]$ based on p_ℓ and p_h are monotonic, and since p_ℓ and p_h are increasing functions of t_ℓ and t_h, respectively, one need only specify the upper and lower extremes of the range on each. Then (2.14) would become

$$\hat{\mathrm{E}}^{+}_{1\alpha}\left[\mathcal{T}(T)\,\middle|\,\mathcal{T}_h, \hat{p}^{-}_{h\overline{\alpha}}\right] \;=\; \left(1-\mathcal{T}_h\right)\hat{p}^{+}_{h\overline{\alpha}} + \mathcal{T}_h \tag{2.56}$$

where h^- is the lower limit of a one-sided confidence interval with subjective probability $1-\alpha$. Because of the additional source of uncertainty, the pseudo-confidence interval has a mixed probability of about $1-2\alpha$. Similarly, one may define the lower bound

$$\hat{\mathrm{E}}^{-}_{1\alpha}\left[\mathcal{T}(T)\,\middle|\,\mathcal{T}_\ell, \hat{p}_{\ell+}\right] \;=\; \mathcal{T}_\ell\,\hat{p}^{+}_{\ell\overline{\alpha}}. \tag{2.57}$$

Theroem 3.5 shows that $\mathcal{T}_1(t) \leq \mathcal{T}_2(t)$ for all $t > 0$ implies $\mathrm{E}[\mathcal{T}_1(T)] \leq \mathrm{E}[\mathcal{T}_2(T)]$. So, if one is using the hybrid approximation, one need know only the upper and lower bounds on $\mathcal{T}(t)$ over its range, especially in (t_ℓ^-, t_h^+). Given estimates of p_ℓ and ζ, one can perform a sensitivity analysis on mean timeliness, using the bounds on $\mathcal{T}(t)$.

Constructing pseudo-confidence intervals

The simplest pseudo-confidence intervals depend upon monotonic relations. For example, increasing the call initiation rate, mean service time, or call-back frequency will increase the load and should lead to lower values of $\mathrm{E}[\mathcal{T}(T)]$. In these cases, one can set up a noise array at the extremes of each of such parameters, one extreme of high, one of lowest load. Ordinarily, combining such independent elements might lead to a very low-probability combination. However, in C^3I systems, external pressures often cause all load factors to move in the same direction.

The pseudo-confidence interval will consist of the extremes of stochastic confidence intervals, determined at the extremes of the ranges in the noise array. This assumes that variance does not increase sharply as one moves away from the extremes. Noting that variance tends to increase with the offered load in queueing systems and that timeliness will tend to be highest at the point of lowest load, this assumption is probably valid at the upper bound of $\mathrm{E}[\mathcal{T}(T)]$. but may not hold at the lower bound. However, since the distribution of noise factors within the range of expected variation is unknown, this adjusted extreme measure is more likely to be accurate than intervals based on the variance induced into $\hat{\mathrm{E}}[\mathcal{T}(T)]$ by the two extremes of load.

An alternative is to select random levels of the noise factors. This is necessary

when the relationship between factors and performance is either unclear or known to be not monotonic. In this case, one defines pseudo-confidence intervals based on the observed variance in the output as a result of the random variation in the noise. This is a weaker method, first because it may be difficult to justify a particular distribution of the noise levels, and second, because this method makes it difficult to pin-point problem spots.

An illustration

The following example will be used to demonstrate some of the points so far. The details of analysis are in Appendix C.

Example 2.1 *The analyst is to select parameters for the simple communications network depicted in Figure 2.14. The nodes A, B, and C represent three information centers that must communicate to complete a given task. The nodes X and Y represent all other users of the network. The message traffic is of a single priority class, and consists of both voice and data. Table 2.2 describes this traffic. Mean call frequency and message length are assumed to be within* $\pm 20\%$ *of the stated values with subjective probability* 95%.

The given task is A, B, and C's role in Activity A. Each trunk channel can support either data at a 9600 baud rate or voice. Although the local networks have enough capacity to be of no concern in this situation, the capacity of each node might. These capacities are listed in Table 2.3

The analyst's choices are summarized in Table 2.4. The analyst must choose between two call-back methods. The first is a conventional one in which the originator calls back after a random wait if the path is busy. The second signals the originator

when the requested call becomes feasible. The analyst must also choose whether to use an eight- or sixteen-channel trunk, and the mix of voice and data channels. The first two choices incur a monetary cost as well as a performance effect. The mix of channel types affects only performance.

The systems will be compared on the basis of their mean system timeliness, $E_S[\mathcal{V}(T)]$, *measured over all needlines participating in the given task. Timeliness will be measured using the* $\mathcal{T}_2(t)$ *model, interpreting perishability codes as in Table A.1 on page 138. The* $\tilde{t}$ *and* t_0 *values are assumed to be within* $\pm 20\%$ *of tabled values with subjective probability* 95%. *Table 2.5 lists the nominal and* 95% *subjective probability limits on importance weights. An acceptable system is one that retains at least* 75% *of its needlines' initial value. The analyst will seek an economical system that yields better or equal* $E_S[\mathcal{V}(T)]$ *than others over the range of expected conditions and effectiveness measures.*

First screening In the screening phase, the analyst tested the system at three, six, and nine each of data and voice trunks. Also, he represented the range of offered load by a 2×2 noise array, representing all four combinations of extreme loads, keeping data and voice load distinct. Figure 2.15 shows p.s.i.s for p_ℓ as a function of the number of data trunks. This implies that the three-trunk options perform poorly. Figure 2.16 shows that high data load has a greater affect on $\hat{p}_\ell$ than high voice load. Finally 2.17 shows the sensitivity of $E_S[\mathcal{T}(T)]$ to the ratio of class importance values. Note that in the case of three data trunks, the effect of the different ratios is not relatively great, but when six or nine data trunks are used, the ratio differences almost double the p.c.i.

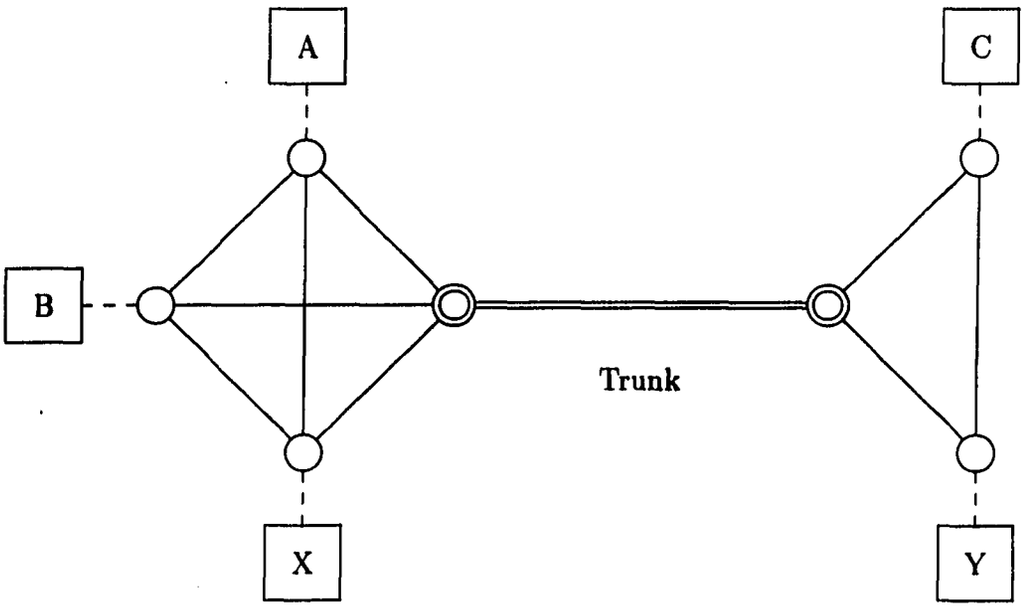


Figure 2.14: Communications Network for Example 2.1

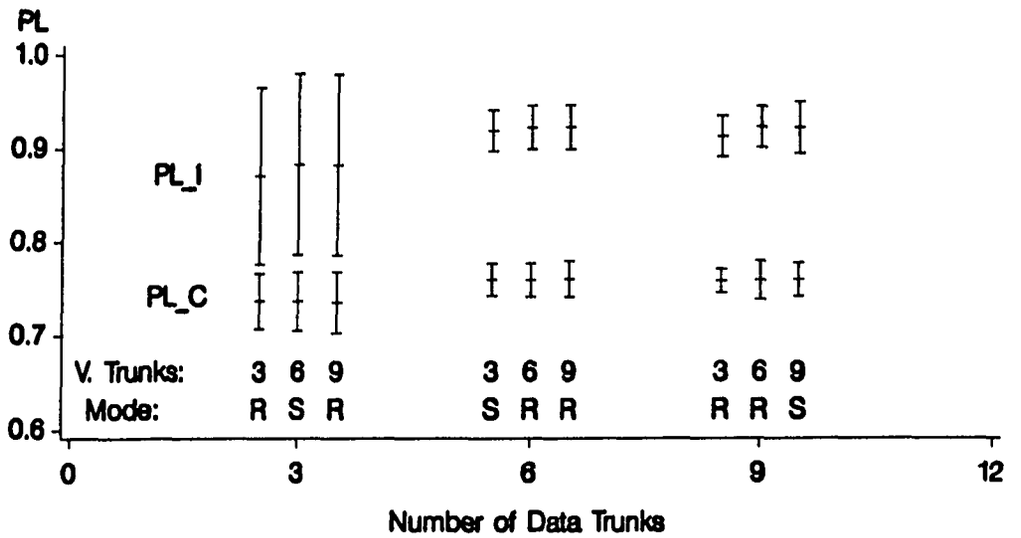


Figure 2.15: 95% Confidence Intervals for the Effect of Network Parameters on p_ℓ During Preliminary Screening

Table 2.2: Needline Specifications for Example 2.1

Need Line	Originator	Receiver	Activity	Mode	Frequency (Calls/Day)	Message Length (kBytes)	Perishability	Cost of Failure
1	A	B	A	V	24	8.0	5	I
2	A	B	A	V	4	72.0	3	I
3	A	B	A	D	960	20.0	6	C
4	A	B	C	D	240	75.0	5	C
5	A	C	A	V	4	32.0	4	I
6	A	C	A	D	960	32.0	6	C
7	A	C	C	D	480	16.0	6	E
8	A	X	A	V	4	48.0	3	I
9	A	X	C	D	4800	4.0	8	C
10	A	Y	C	D	2400	8.0	7	C
11	B	A	A	V	24	0.4	5	I
12	B	C	A	V	48	2.0	5	I
13	B	C	A	D	1440	4.0	6	I
14	B	C	C	D	720	12.0	5	C
15	C	B	A	V	48	1.6	6	I
16	C	B	C	D	1200	4.0	6	C
17	C	A	A	D	480	8.0	5	C
18	X	A	A	V	12	2.0	4	C
19	X	Y	A	V	480	4.0	8	I
20	X	Y	A	D	9600	4.0	9	I

Table 2.3: Node Characteristics in Example 2.1

Node	Capacity	On Task?
A	10	Yes
B	5	Yes
C	5	Yes
X	20	No
Y	20	No

Table 2.4: Network Options and Monetary Costs in Example 2.1

Option	Call-back System	Trunk Capacity	Relative Cost
1	Random	8	1.00
2	Scheduled	8	1.05
3	Random	16	1.50
4	Scheduled	16	1.55

Table 2.5: Cost of Failure Values for Example 2.1

Cost of Failure		Value		
Code	Description	Minimum	Nominal	Maximum
I	Indispensable	2.0	4.0	16.0
C	Critical	1.4	2.0	4.0
E	Essential	1.0	1.0	1.0

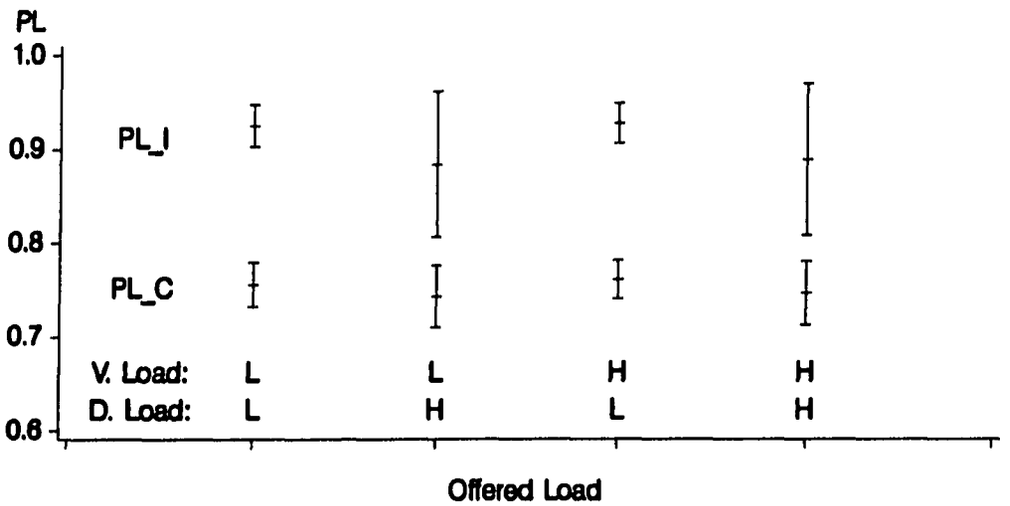


Figure 2.16: 95% Confidence Intervals for the Effect of the Offered Load on p_ℓ During Preliminary Screening

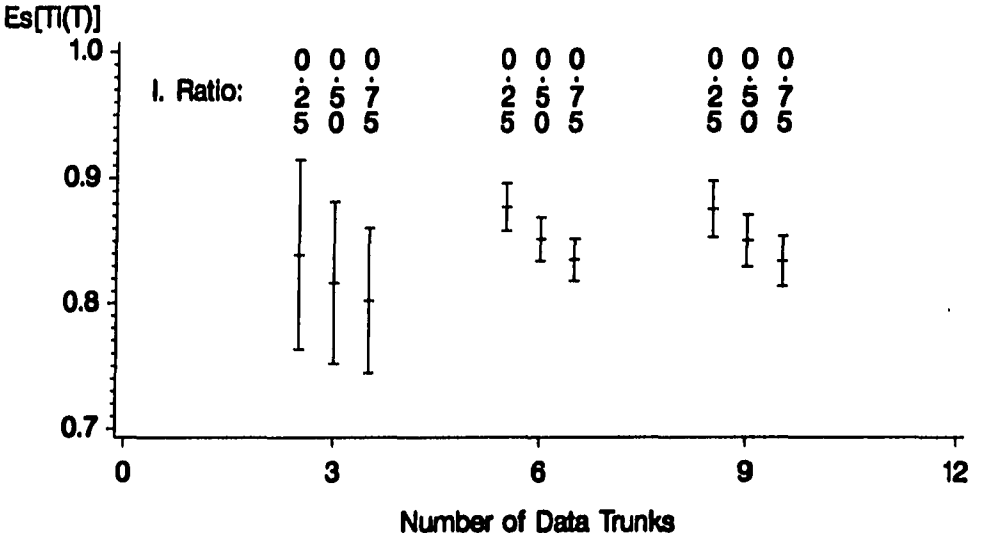


Figure 2.17: 95% Confidence Intervals for the Effects of the Importance Ratio and Number of Data Trunks on $E_S[\mathcal{T}(T)]$ During Preliminary Screening

In spite of the uncertainty in the C^2 load and $\mathcal{V}(t)$, the analyst can conclude

1. More than three data and voice trunks will be needed,
2. The eight-trunk option will probably work,
3. The variance in the data load affects timeliness greatly, and
4. The data and voice loads interact.

Second screening In the second set of runs, the analyst decided to test the eight-trunk option, but kept data and voice load distinct. In addition, the analyst sought further C^2 information on four high-traffic needlines. Table 2.6 shows the revised C^2 information at this stage. Finally, he included a "baseline" model, in which the number of trunks was not a constraint, for comparison.

Table 2.6: Results of Investigating the Range of Load Factors

Needline	Range Factors for θ_a and θ_s	
	New Lower Limits	New Upper Limits
6	0.80	1.20
10	0.90	1.10
19	0.95	1.05
20	0.95	1.05

Figure 2.18 illustrates pseudo-confidence intervals for $E_S{}^-$. The symbol "99" represents the baseline model. It appears that the eight-trunk option is satisfactory and that any of 4, 5, or 6 data trunks will perform equally-well. It also appears that the two call-back modes perform equally-well. Furthermore, Figures 2.19 and 2.20 show that uncertainties in I values and the exact values of t_ℓ and t_h have much larger effects than those observed in the analysis, indicating that further analysis without tightening up these functions is pointless.

Problems of statistical coverage

Of course, intervals with subjective probabilities present problems of validity and acceptance. However, even if the individual C^2 intervals are accepted, the resulting C^3I intervals may be difficult to work with. For example, The CDB specifies arrival rates and message lengths. If these values are stated to within $\pm 10\%$ with subjective probability 0.95, then the same can be said of needline mean interarrival times and mean service times. That is, if subjective probability is 0.95 that there are between 18 and 22 calls per hour, the mean interarrival time is between 2.73 and 3.33 minutes with the same subjective probability. However, some C^3I factors may depend upon multiple C^2 factors. For example, suppose θ_r, the mean time between recall attempts

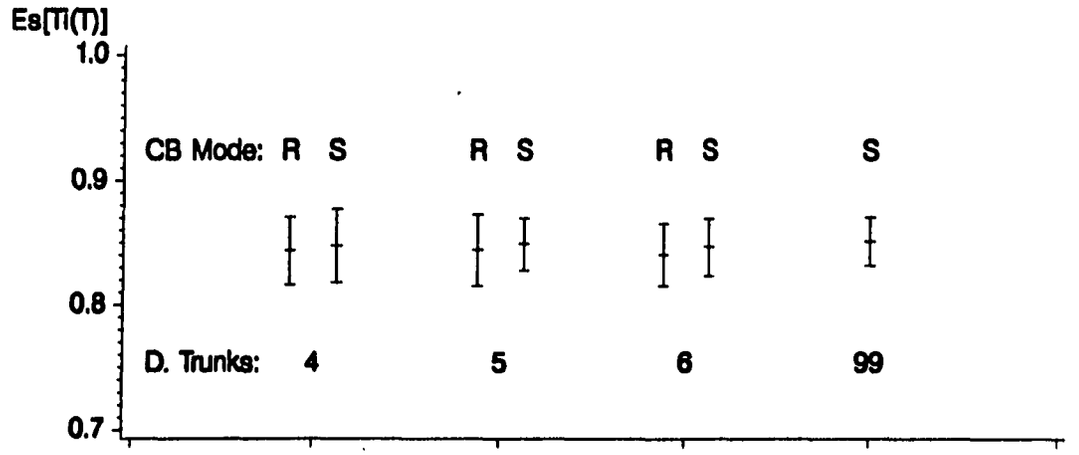


Figure 2.18: 95% Confidence Intervals for the Effect of System Parameters on $\hat{E}_S$ From the Second Screening

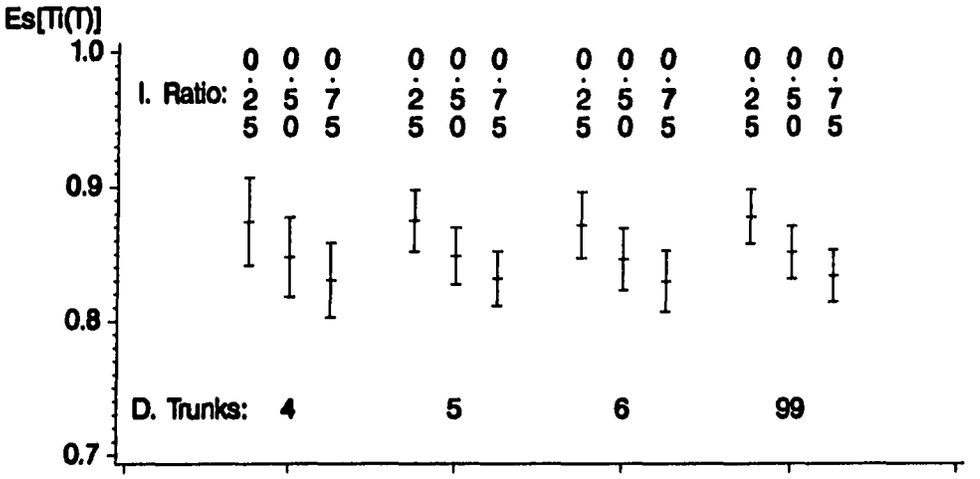


Figure 2.19: 95% Confidence Intervals for the Effect of Timeliness Functions over the Range of Timeliness Models and System Parameters on $\hat{E}_S$ From the Second Experiment

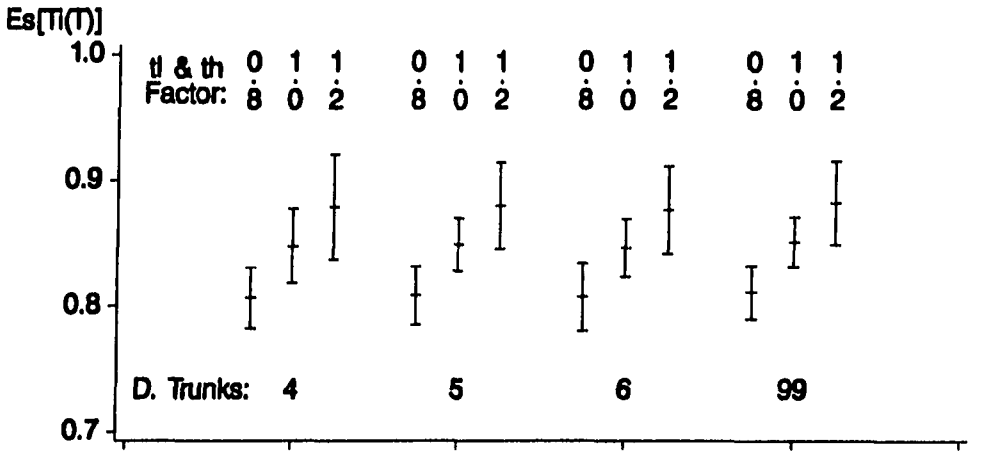


Figure 2.20: 95% Confidence Intervals for the Effect of Importance Ratios over the Range of Timeliness Models and System Parameters on $\hat{E}_S$ From the Second Experiment

is $\theta_a U/I$, where θ_a is the initial call frequency, U the urgency of the call, and I the importance of the call. If each can vary $\pm 10\%$ with subjective probability 0.95, then θ_r can vary from -26% to $+34\%$ with subjective probability about 0.85. The result can be very wide intervals of low confidence.

Using the Proposed Devices to Control Analysis

The following describes the a basic plan, first from the perspective of overall three-phase problem-solving process, then as a repetitive process within each phase. These two views are illustrated in Figures 2.21 and 2.22. It also compares this approach to that of the MCES and Taguchi.

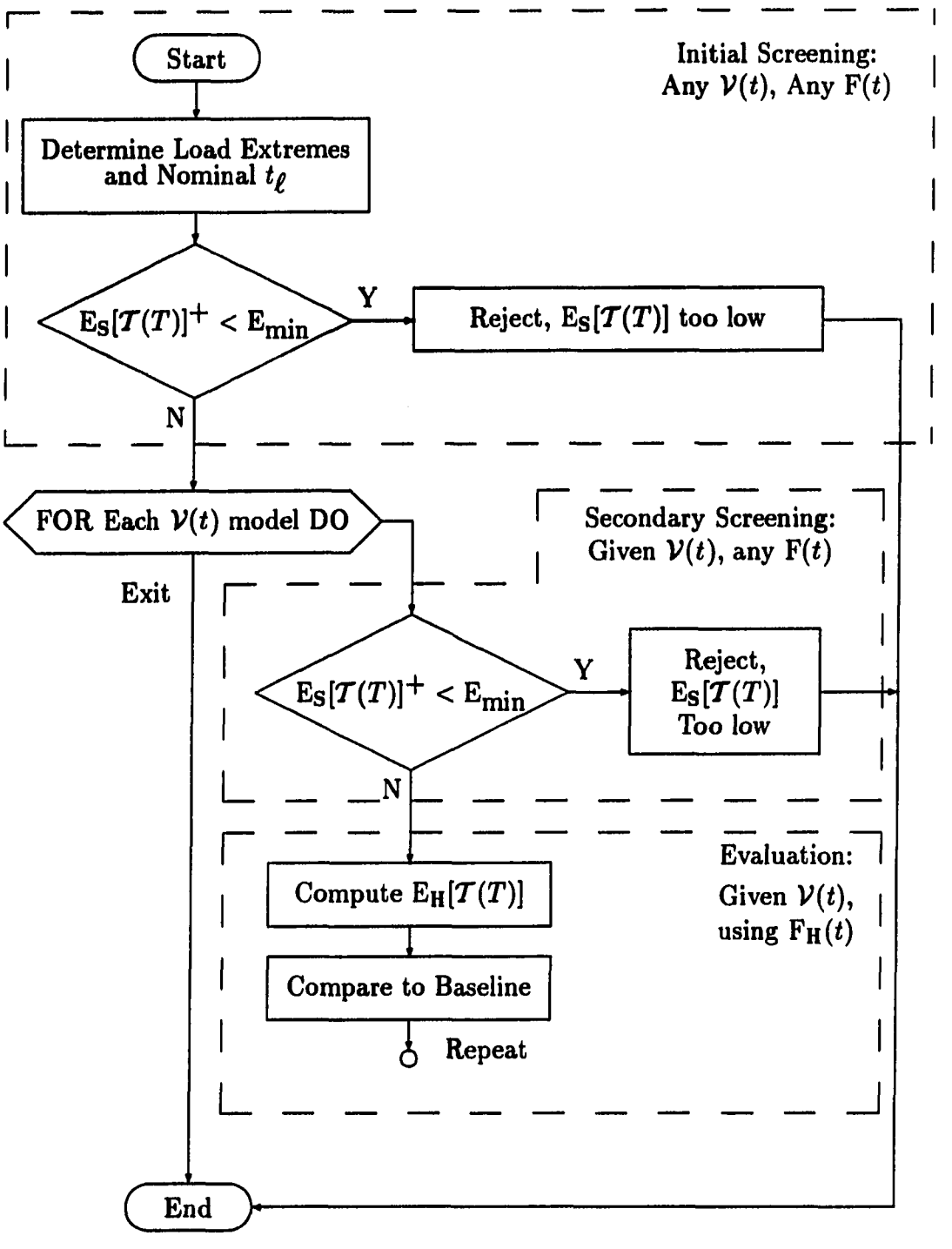


Figure 2.21: Overview of the Proposed Method

Overall plan

The method involves a sequence of iterative cycles which 1) make closer and closer approximations to either $E_S[\mathcal{T}(T)]$ or $\min_S[\mathcal{T}(T)\,|I]$, 2) make more and more assumptions about $F(t)$ and $\mathcal{T}(t)$, and 3) are more costly. The goal is to arrive at results that are as acceptable and economical as possible, given the state of the C^2 data. C^3I system analysis usually consists of three distinct phases: modeling, screening, and evaluation. Figure 2.21 illustrates the overall plan for the last two stages. Although estimates are rough in the initial screening, they apply to all $\mathcal{V}(t)$ and $F(t)$ models. In the secondary screening, one tests results for a range of $\mathcal{V}(t)$ models, but still makes no assumptions about $F(t)$. In the final stage, analysts must make assumptions about both $\mathcal{V}(t)$ and $F(t)$. If all goes well, one ends up with at least one good system, but if not, one can at least identify the reason none is found.

Procedure details

Figure 2.22 depicts the events that occur within each phase, including modeling, in greater detail than in Figure 2.21. Analysts repeat this loop until the phase problem is resolved. At the end of each repetition, they ask "What can we possibly gain from further analysis?" Depending on the answer, they will refine and repeat analysis within the current phase, continue to the next phase, return the problem to the C^2 analysts for further refinement, or terminate the analysis.

The procedure in Figure 2.22 begins with a *formulation of the problem.* *C^2 data* consists of specifications, such as offered load, the scenario, and effectiveness measures, in C^3I-system-independent form. The *C^3I data* consists of equipment models and possibly data from earlier C^3I analyses. The C^3I analysts reformulate this

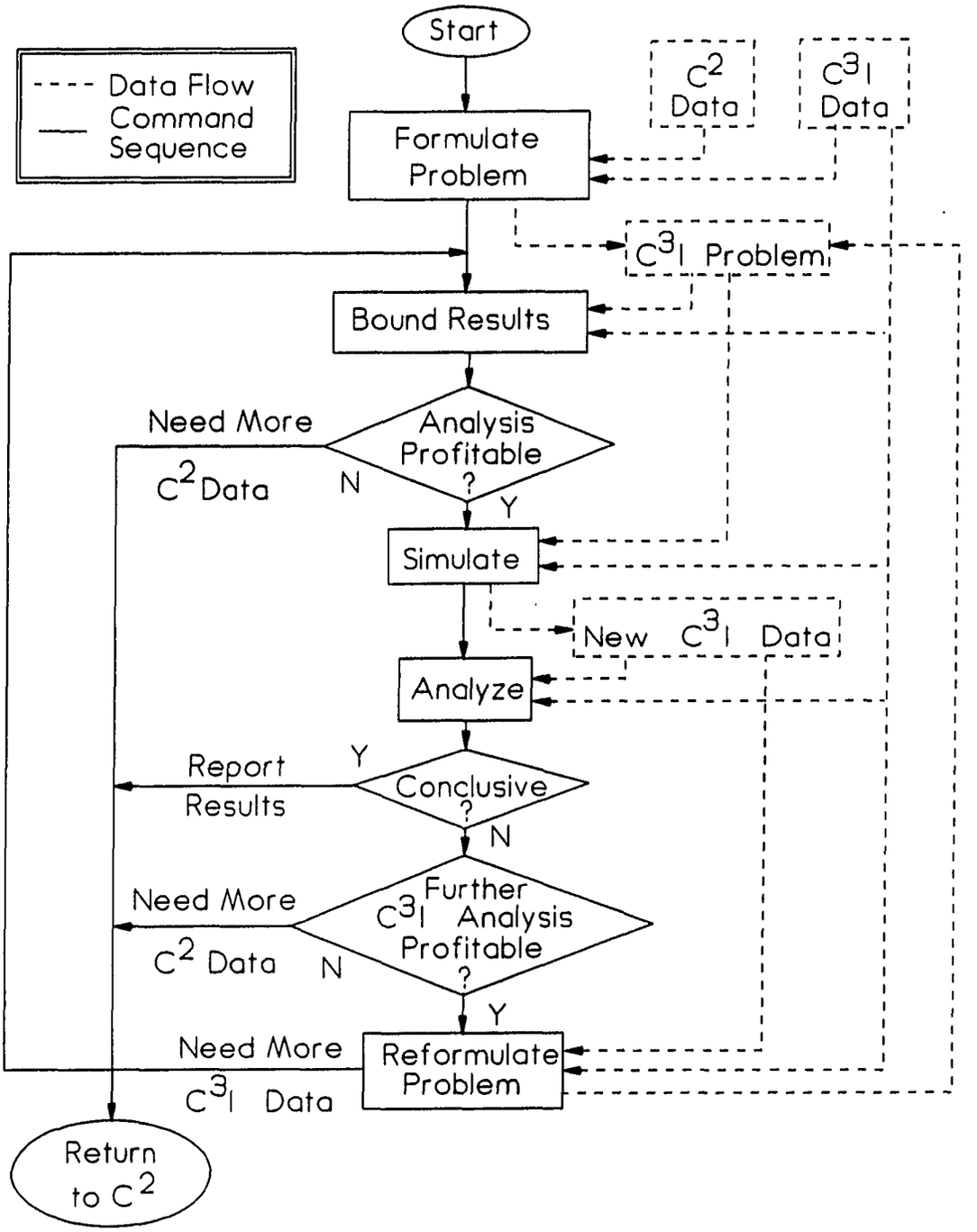


Figure 2.22: General Plan of Each Cycle of the Proposed Method

data into a self-contained C^3I simulation model and scenarios, collectively referred to as the *C^3I problem.*

Because simulations take time and money to develop and run and because analysis is expensive, the C^3I analysts next attempt to bound the results. The goal is to determine if the problem is sufficiently well-stated to yield clear results. All data is given in the form of intervals. This step checks that these intervals are narrow enough to allow meaningful output intervals. If this is not true, the C^3I analysts will need to return the problem to the C^2 experts for further refinement.

The *action* step will generate new C^3I data. The C^3I analysts then *analyze* this new data, together with the original C^2 data, to attempt to find results. If they find conclusive results, such as "System A is the only one that provides an acceptable grade of service," they report them to the C^2 experts. If results are not conclusive, the C^3I experts determine whether or not further C^3I analysis could be profitable. If so, they reformulate the problem and either repeat the current phase or go on to the next. If further C^3I analysis would not be profitable, they return the problem to the C^2 experts for further refinement.

Comparison of the proposed method to the MCES

The method is similar to the MCES approach in that the C^3I analysts estimate MOPs, which the C^2 analysts convert into effectiveness measures. This separation of functions assures that each task is undertaken by those most qualified to handle it. However, the method goes beyond the MCES by using FMOPs to increase the amount of information passed between the C^2 and C^3I community. This improves concurrency and helps the C^3I experts concentrate on types of analysis that will be

most useful to the C^2 experts.

Comparison of the proposed method to Taguchi's method

The method is similar to Taguchi's method in that it utilizes factorial design and exercises the system over a range of expected operating conditions. It also considers all operational conditions and system characteristics intervals, rather than point values. However, unlike Taguchi's method, it continually asks the value of further analysis and considers the possibility of returning a problem to the C^2 group. In this way, it attempts to find the most profitable areas to work in, whether they be problem description, simulation and analysis, or refinement of effectiveness measures. Finally, while Taguchi's method uses a well-defined loss function, the proposed method considers the loss function poorly defined. This extension allows one to work when quality is only approximately defined. The use of FMOPs allows one to measure the sensitivity of results to changes in the loss function without additional simulation.

Adaptations to Steady-State Analysis in Autoregressive Systems

The analysis above assumes the observations are independent. If one is observing dedicated C^3I systems or making a transient analysis, this is a reasonable assumption. However, if several needlines share common resources and one wishes steady-state averages of system effectiveness, then this is no longer a viable assumption. Output sequences of task completion times in networks tend to be highly correlated, especially if the system load is high.

Basic method

Fortunately, all parameter estimates above are based on probabilities and means. This avoids most of the complications due to autocorrelation. The chief problem is how to adjust for the artificially low estimates of estimator variance. The basic procedure is similar to the common "method of batch means," with a few variations. Its main steps are:

1. Truncate the initial part of the output record to minimize the effect of initial conditions and possible covariance between independently-seeded runs,
2. Estimate the lag correlation ($\mathcal{L}$) and form batches of length $5\widehat{\mathcal{L}}$,
3. Estimate the variance of the estimator either as if the batch means are approximately normally-distributed, or based on other assumed properties of the estimator, and
4. Estimate confidence limits on bounds for $E_S[\mathcal{T}(T)]$ either from the bounds of the basic estimators or as functions of estimators with assumed properties.

The following describes adaptation of the devices above to this method of coping with autocorrelation. Although parts of this approach are open to criticism, these are criticisms of the basic "method of batch means" rather than the devices proposed above. That is, the proposed devices may not make this method work better, but at least they do not make matters worse.

Initial truncation

A key assumption is that enough of the beginning of the trace has been deleted that the remaining trace is essentially a stationary random process. This is probably the most difficult part of the method, since there are no universally-approved truncation tests [39, 40, 41]. The method used in the examples is to

1. Estimate critical lag ($\mathcal{L}$) through tables of autocorrelation [42]. Call this estimate L_0.

2. examine a trace of non-overlapping batch means of about $L_0/5$ observations each from five independently-seeded runs of about $100L_0$ in length. Let N_0 be the point at which the initial bias and correlation between runs appears to fade.

3. Test the adequacy of N_0, using Schruben's F-ratio test (See Appendix B for details). If too small, let $N_1 = 2N_0$ and test N_1. Repeat, extending the runs if necessary.

Point estimates

If the process $\{T_i\}$ is stationary, then the T_i are identically, distributed, but dependent. Because the relation $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$ holds whether X and Y are correlated or not, the average of successive observations from a stationary random process is simply $\mathrm{E}[T]$. Estimates of p_x are means of the process $\{X_i\}$, where $X_i = 1$ if $T_i < t_x$, and zero otherwise. If $\{T_i\}$ is stationary, then so is $\{X_i\}$, and the same principle applies.

It is important to note that the estimates are of the distribution of the stationary process, not its elements. Let $T_i = S_i + \sum_{j=1}^{\infty} G_j(S_{i-j})$ represent the components of the process. The S_i represent the independent element of the process and the $G_j(S_{i-j})$ represent the effect of autocorrelation on the T_i. Because the process is stationary, $\mathrm{E}[\sum_{j=1}^{\infty} G_j(S_{i-j})] = 0$ and $\mathrm{E}[T] = \mathrm{E}[S]$. However, the effect of $\sum_{j=1}^{\infty} G_j(S_{i-j})$ is to distort the shape of the distribution. For example, in a $(M/M/1)$ queue, although the distribution of S is exponential with mean $1/\mathrm{E}[T]$, the autocorrelation will make larger time values more probable in the distribution of T than in that of S.

This is not a problem in assessing C^3I performance because it is the distribution of T is of interest, rather than that of S. It is T that affects performance, not S, since the shift in probabilities will also cause a shift in expected timeliness. Estimators of $\mathrm{E}[\mathcal{T}(T)]$ based on probability estimates of T are unbiased since they are based on actual counts of T_i in selected intervals and make no assumptions about the distribution of T. The hybrid estimator, based on ζ will be a biased estimator of $\mathrm{E}[\mathcal{T}(T)]$ because of its assumption of exponential times. However, as shown in Chapter 3, the bias due to the shape of the distribution would be very small.

Estimator variance

The remaining problem is that of the low variance of the T_i. As shown in Appendix B, each observation provides only a fraction of the information an independent observation would. The approach is the "method of batch means" is to estimate the lag correlation $\mathcal{L}$ in the output record, then compute the means of adjacent, but non-overlapping batches of size $5\hat{\mathcal{L}}$. The $n/5\hat{\mathcal{L}}$ batch means are then treated as

uncorrelated, normally-distributed observations.

A variation is proposed to estimate probabilities. It uses the presumed structure of the estimator, but adjusts the apparent sample size. Thus,

$$\widehat{\mathrm{Var}}(\hat{p}_x) = \frac{\hat{p}_x(1-\hat{p}_x)}{n/(5\hat{\mathcal{L}})} \tag{2.58}$$

$$\widehat{\mathrm{Var}}(\hat{p}_h) = \frac{\hat{p}_h\left(1-\hat{p}_h\right)}{n/(5\hat{\mathcal{L}})} \tag{2.59}$$

$$\widehat{\mathrm{Cov}}(\hat{p}_\ell, \hat{p}_h) = \frac{\hat{p}_\ell\left(1-\hat{p}_h\right)}{n/(5\hat{\mathcal{L}})} \tag{2.60}$$

where n is the number of observations in the steady-state portion of the output. This uses the information about the distribution of p_x, which the basic method does not. In addition, since one need not actually calculate the batch means, it is somewhat more efficient. $\mathrm{Var}(\hat{\zeta})$ was estimated from the variance among the $n/(5\hat{\mathcal{L}})$ individual batch means, which is the usual way in this method.

Another variation is to simply not batch [43]. It is difficult to tell in advance exactly how may observations to generate. One can estimate the variance of $\hat{p}$ from its value and the number of independent runs. A single high-intensity segment can greatly increase $\hat{\mathcal{L}}$. In one batch of 120 independent simulated 24-hour periods, this researcher got $\mathcal{L}$ estimates ranging from 20 to 200. If $\hat{\mathcal{L}}$ is too large, the output is too highly averaged, if too small, observations are too dependent. Either will reduce variance estimates. In addition, the differing $\hat{\mathcal{L}}$ estimates produce differing numbers of observations per run, complicating ANOVA.

One needs at least five replicates to detect initial bias. In some cases, this may be enough observations to provide good power. Another option is to select $\hat{\mathcal{L}} = \max_{j=1}^{m} \hat{\mathcal{L}}_j$, where $\hat{\mathcal{L}}_j$ is the estimator for the j-th run. This argues that lag

correlation is a system, rather than a run, characteristic. Table 2.7 is the result of selecting $\hat{\mathcal{L}}$ so that each run generated 20 observations. The tables entries are estimates based on the 100 observations total in five replicates. The standard deviation estimates are remarkably consistent, indicating promise to this approach. However, the estimate of variance within each run was less than that between runs, indicating some problem. One possibility is that the selected $\mathcal{L}$ was too large. Another is that some higher-order correlation is at work, which is too subtle to be detected, but significant, meaning $\mathcal{L}$ is too small.

Summary

Two key devices are proposed. The first is the use of focused measures of performance to improve communications between C^2 and C^3I regarding regions of greatest importance to system effectiveness measures. The second is the use of pseudo-confidence intervals to transfer what is not known about the C^2 problem to the C^3I analyst and back again to the C^2 expert. It is shown that one can differentiate among C^3I systems with only partial C^2 information, but cannot determine the quality of C^3I analysis without knowing the quality of the C^2 problem formulation.

These devices, used with a strategy of balancing necessary assumptions with needed information can not only make C^3I analysis a more reliable predictor of actual C^3I system behavior, but can also identify C^2 parameters that have the greatest effect on the resolution of the C^3I analysis. In general, they direct C^2 and C^3I specialists into mutually-beneficial areas of analysis.

These devices expose some problems and leave others unresolved. The problem of subjective probabilities is very difficult to resolve, but as noted above, exists

Table 2.7: Results of the Initial Screening

Obs	Call Back Mode	Voice Trunks	Data Trunks	Voice Load	Data Load	Observed Values of P_ℓ Max	Min	Mean	S
1	R	3	3	-1	-1	0.933	0.688	0.814	0.044
2	R	3	3	-1	1	0.853	0.600	0.763	0.043
3	R	3	3	1	-1	0.894	0.705	0.812	0.043
4	R	3	3	1	1	0.849	0.669	0.768	0.037
5	R	3	9	-1	-1	0.937	0.674	0.824	0.047
6	R	3	9	-1	1	0.900	0.725	0.815	0.029
7	R	3	9	1	-1	0.900	0.719	0.825	0.039
8	R	3	9	1	1	0.865	0.665	0.814	0.028
9	R	6	6	-1	-1	0.931	0.697	0.824	0.044
10	R	6	6	-1	1	0.875	0.706	0.815	0.028
11	R	6	6	1	-1	0.931	0.731	0.833	0.042
12	R	6	6	1	1	0.869	0.661	0.820	0.031
13	R	6	9	-1	-1	0.937	0.684	0.825	0.045
14	R	6	9	-1	1	0.867	0.706	0.814	0.029
15	R	6	9	1	-1	0.931	0.747	0.835	0.041
16	R	6	9	1	1	0.902	0.702	0.823	0.026
17	R	9	3	-1	-1	0.918	0.665	0.816	0.046
18	R	9	3	-1	1	0.854	0.624	0.765	0.043
19	R	9	3	1	-1	0.920	0.714	0.823	0.045
20	R	9	3	1	1	0.853	0.642	0.769	0.042
21	R	9	6	-1	-1	0.931	0.697	0.825	0.043
22	R	9	6	-1	1	0.875	0.706	0.816	0.026
23	R	9	6	1	-1	0.931	0.747	0.835	0.041
24	R	9	6	1	1	0.876	0.711	0.821	0.025
25	S	3	6	-1	-1	0.931	0.697	0.824	0.044
26	S	3	6	-1	1	0.873	0.698	0.817	0.028
27	S	3	6	1	-1	0.931	0.739	0.831	0.040
28	S	3	6	1	1	0.865	0.706	0.818	0.024
29	S	6	3	-1	-1	0.933	0.688	0.815	0.045
30	S	6	3	-1	1	0.883	0.638	0.767	0.042
31	S	6	3	1	-1	0.911	0.735	0.824	0.043
32	S	6	3	1	1	0.853	0.629	0.770	0.045
33	S	9	9	-1	-1	0.937	0.684	0.825	0.045
34	S	9	9	-1	1	0.882	0.727	0.812	0.027
35	S	9	9	1	-1	0.931	0.747	0.834	0.040
36	S	9	9	1	1	0.902	0.767	0.825	0.023

whether we choose to address it or not. As in other forms of applied analysis, this method leaves open questions of independence, distribution, and effective sample size. Because the need for analysis implies a lack of needed information, the best we can do is use what seems to work well in similar cases and check our results. At least, since the method relies on simple mean estimates and no more than two parameters, it keeps the level of complexity low.

The remaining loose end is the robustness of the hybrid estimator. The claim is that it is not sensitive to the types of distributions common in C^3I analysis. If this claim is true, than $E_H[\mathcal{T}(T)]$ is also a robust estimator of mean timeliness. This point is explored in Chapter 3.

CHAPTER 3. ROBUSTNESS OF THE HYBRID APPROXIMATION

Introduction

Chapter 3 examines the robustness of the hybrid approximation to $E[\mathcal{T}(T)]$ for a single needline. Robustness is be measured by the distribution of estimation bias as a function of the characteristics of the true distribution over a wide range of circumstances. A robust estimator is one for which the estimation bias is small and relatively insensitive to the shape of $F(t)$.

After defining key terms, the following discusses normalized distributions, and using that concept, briefly examines selected features of non-negative random variables and establishes some bounds on $F(t)$. It also examines generalizes phase branching processes. It then extends these results to timeliness. Finally, it discusses the bias of the hybrid approximations in general and within the context of the proposed method.

This analysis shows that the hybrid approximation is much more robust than conventional one- or two-parameter models of $F(t)$. In addition, it determines limits on bias in the case of the $\mathcal{T}_2(t;\tilde{t},t_0)$ model, in general and using the CDB data.

Notation

Approximations of mean timeliness

Let $\mathrm{E}_{\mathrm{X}}[\mathcal{T}_{\mathrm{Y}}(T;\Upsilon)|\,\Phi_{\mathrm{X}}]$ denotes the mean timeliness, assuming T is distributed as a X random variable, with parameter set Φ_{X}, using the Y model of timeliness with parameter set Υ. For example,

$$\begin{aligned}\mathrm{E}_{\mathrm{M}}\left[\mathcal{T}_2(T;\tilde{t},t_0)\middle|\,\theta\right] &= \int_0^{t_0}\frac{1}{\theta}\exp\left(-\frac{t}{\theta}\right)+\int_{t_0}^{\infty}\exp\left(-\frac{t-t_0}{\tilde{t}}\right)\frac{1}{\theta}\exp\left(-\frac{t}{\theta}\right)\,dt \\ &= 1-\left(\frac{\theta}{\tilde{t}+\theta}\right)\exp\left(-\frac{t_0}{\tilde{t}}\right) \qquad (3.1)\end{aligned}$$

is the mean of the $\mathcal{T}_2(t;\tilde{t},t_0)$ timeliness model when T is distributed as an exponential random variable with mean θ. Here $\Phi_{\mathrm{X}}=(\theta)$ and $\Upsilon=(\tilde{t},t_0)$.

If a rule or statement is meant to apply generally, the subscript or parameter sets will not be stated. For example, $0\le \mathrm{E}[\mathcal{T}(T)]\le 1$ is true for all distributions, timeliness models, and parameter sets.

Bias

The bias is the signed difference of the (assumed) true mean timeliness less the approximate mean timeliness[1]. Subscripts indicate both distributions and the timeliness model. In general,

$$\begin{aligned}&\mathcal{B}_{\mathrm{X}|\mathrm{Z}}[\mathcal{T}_{\mathrm{Y}}(T;\Upsilon)|\,\Phi_{\mathrm{X}}\cup\Phi_{\mathrm{Z}}] \\ &\qquad = \mathrm{E}_{\mathrm{Z}}[\mathcal{T}_{\mathrm{Y}}(T;\Upsilon)\,|\Phi_{\mathrm{Z}}]-\mathrm{E}_{\mathrm{X}}[\mathcal{T}_{\mathrm{Y}}(T;\Upsilon)\,|\Phi_{\mathrm{X}}] \qquad (3.2)\end{aligned}$$

[1]This definition of bias is the negative of the usual definition. This change was made to make graphs easier to interpret.

is the bias of estimating $\mathrm{E}[\mathcal{T}_{\mathrm{Y}}(T)]$ using the $\mathrm{F}_{\mathrm{X}}(t)$ distribution when the true distribution is F_{Z}. For example,

$$\mathcal{B}_{\mathrm{M}|\Gamma}\left[\mathcal{T}_2(T;\tilde{t},t_0)\middle|\,\theta,\varphi\right] \;=\; \mathrm{E}_{\Gamma}\left[\mathcal{T}_2\left(T;\tilde{t},t_0\right)\middle|\,\theta,\varphi\right] \;-\; \mathrm{E}_{\mathrm{M}}\left[\mathcal{T}_2\left(T;\tilde{t},t_0\right)\middle|\,\theta\right]$$

is the bias when using the exponential distribution to estimate $\mathrm{E}[\mathcal{T}_2(T)]$ when T is a gamma random variable. As before, subscripts and parameter sets will be omitted if a statement is meant to apply to all distributions, timeliness functions, or parameter sets.

One may either compute the bias directly from (3.2), or integrate by parts to yield:

$$\mathcal{B}_{\mathrm{X}|\mathrm{Z}}\left[\mathcal{T}_{\mathrm{Y}}(T)\middle|\,\Phi_{\mathrm{X}}\cup\Phi_{\mathrm{Z}}\right] \;=\; -\int_0^{\infty}\Delta_{\mathrm{X}|\mathrm{Z}}\left(t;\Phi_{\mathrm{X}}\cup\Phi_{\mathrm{Z}}\right)d\mathcal{T}_{\mathrm{Y}}(t) \tag{3.3}$$

where

$$\Delta_{\mathrm{X}|\mathrm{Z}}\left(t;\Phi_{\mathrm{X}}\cup\Phi_{\mathrm{Z}}\right) \;=\; \mathrm{F}_{\mathrm{Z}}\left(t;|\Phi_{\mathrm{Z}}\right)-\mathrm{F}_{\mathrm{X}}\left(t;|\Phi_{\mathrm{X}}\right). \tag{3.4}$$

This form is quite useful, since $\mathrm{F}(t)$ is better behaved than $\mathrm{f}(t)$ and the derivative of timeliness function is generally known.

The Markovian approximation

The *Markovian approximation* is an estimator of $\mathrm{E}[\mathcal{T}(T)]$ which assumes $\mathrm{F}(t)$ is exponential with mean $\theta = \mathrm{E}[T]$. Since it considers only $\mathrm{E}[T]$, its use is equivalent to comparing systems on the basis of their mean time. It is important in some derivations below. It also reflects the common practice of using the exponential distribution as a first approximation. In general,

$$\hat{\mathrm{E}}_{\mathrm{M}}\left[\mathcal{T}\left(T;\Upsilon\right)\middle|\,\hat{\theta}\right] \;=\; \int_0^{\infty}\mathcal{T}\left(t;\Upsilon\right)\left(\frac{1}{\hat{\theta}}\right)\exp\left(-\frac{t}{\hat{\theta}}\right)dt. \tag{3.5}$$

Since $\hat{\mathrm{E}}_{\mathrm{M}}[\mathcal{T}(T)]$ is independent of all shape parameters, it is bounded by the smallest and largest possible values of $\mathrm{E}[\mathcal{T}(T)|\hat{\theta}]$. Thus, the bias can be either positive or negative and of great magnitude, especially if φ is large.

Expected FMOP values

The comparisons made in this chapter are based on the expected values of FMOPs which estimate model parameters. The expected value of p_x is simply $\mathrm{F_T}(t_x)$. If T is exponentially distributed with mean θ, then because of the Markovian property, $\zeta = \theta$. Otherwise, let $\{\grave{t}\}$ be a sequence formed from only those $t_i > t_\ell$. For those t_i, $\grave{t}_i = t_i - t_\ell$, with obvious renumbering. Furthermore, let r be the number of $\grave{t} < t_h - t_\ell$. Then if $\hat{\zeta}_{\grave{n}}$ is an estimate of ζ based on $\grave{n}$ tasks, then

$$\begin{aligned}\hat{\zeta}_{\grave{n}} &= \left(\frac{1}{r}\right)\sum_{i=1}^{\grave{n}} \min\left(\grave{t}_i, t_h - t_\ell\right) \\ &= \left(\frac{1}{r}\right)\left[\left(\sum_{\grave{t}_i < t_h - t_\ell} \grave{t}_i\right) + (\grave{n} - r)(t_h - t_\ell)\right].\end{aligned}$$

Fixing $r/\grave{n}$ and taking the limit as $\grave{n} \to \infty$ leads to:

$$\begin{aligned}\mathrm{E}[\hat{\zeta}; t_\ell, t_h|\mathrm{F}(t)] &= \mathrm{E}[T\,|t_\ell \le T < t_h] - t_\ell + \left(\frac{1-p_h}{p_h - p_\ell}\right)\left(t_h - t_\ell\right) \\ &= \left(\frac{1}{p_h - p_\ell}\right)\left[\int_{t_\ell}^{t_h} t\, d\mathrm{F}(t) + t_h\left(1 - p_h\right) - t_\ell\left(1 - p_\ell\right)\right] \\ &= \frac{\int_{t_\ell}^{t_h} d\mathrm{S}(t)}{p_h - p_\ell} \qquad (3.6)\end{aligned}$$

where $\mathrm{S}(t) = 1 - \mathrm{F}(t)$ is the survivor function of T. So, in the two-hop example,

$$\mathrm{E}[\hat{\zeta}; t_\ell, t_h|\mathrm{F}(t)] =$$

$$\left(\frac{1}{p_h - p_\ell}\right)\left\{1 + \frac{\exp\left[-(\mu_2 - \lambda)t_\ell\right] - \exp\left[-(\mu_2 - \lambda)t_h\right]}{\mu_2 - \lambda}\right\}. \qquad (3.7)$$

Note than if $t_\ell = 0$ and $t_h \to \infty$, (3.7) reduces to $\mathrm{E}[\hat{\zeta}] = 2/(\mu_2 - \lambda)$, as expected.

The Distribution of Non-Negative Random Variables

Suppose an analyst wishes to characterize some non-negative random variable, T, on the basis of simulation output. In order to use simulation, the analyst would have to be reasonably certain that T has finite moments, so it is safe to assume that $\theta = \mathrm{E}[T]$ and $\sigma^2 = \mathrm{Var}[T]$ exist. Beyond that, the analyst would be unsure of the nature of $\mathrm{F_T}(t)$.

This section comments on pertinent properties of non-negative random variables with finite moments. It begins by defining *normalized distributions*, which are analogous to *standardized distributions* of unrestricted random variables. It then illustrates common features of several distributions and presents several bounds on $\mathrm{F_T}(t)$. Finally, the discussion turns to the special properties of low- and high- variance distributions. Although this analysis differs from most by making very few assumptions about T, it is sufficient to bound estimation bias, when combined with the concept of focused measures of performance and the proposed method of estimating $\mathrm{E}[\mathcal{T}(T)]$.

Normalized distributions

The *normalized distribution*, sometimes called a scaled distribution, is similar to a standardized distribution, but differs in several important ways. First, the mean of a unrestricted random variable is a location parameter. However, because the mean of a non-negative random variable indicates the distance between the origin and the

expected value, it is a scale parameter. Second, there is no location parameter[2]. Third, all other parameters deal with shape.

For example, a conventional representation of the gamma distribution function, $F_\Gamma(t)$, is:

$$\begin{aligned} F_\Gamma(t;\alpha,\beta) &= \int_0^t \frac{t^{\alpha-1}\exp\left(-\frac{t}{\beta}\right)}{\beta^\alpha \Gamma(\alpha)} dt \\ &= g\left(\frac{t}{\beta};\alpha\right) \end{aligned} \tag{3.8}$$

where

$$\Gamma(\alpha) \triangleq \int_0^\infty x^{\alpha-1}\exp(-x)dx, \quad \alpha > 0. \tag{3.9}$$

is the *gamma* function, and

$$g(t;\alpha) \triangleq \int_0^t \frac{x^{\alpha-1}\exp(-x)}{\Gamma(\alpha)} dx \tag{3.10}$$

is the *incomplete gamma function* [44].

Let θ be the mean of T and $\varphi = \sqrt{\mathrm{Var}(T)}/\theta$ the coefficient of variation. Since $\theta = \alpha\beta$ and $\varphi = \sqrt{1/\alpha}$, the normalized form is:

$$F_\Gamma(t;\theta,\varphi) = g\left(\frac{t}{\theta\varphi^2};\frac{1}{\varphi^2}\right) \quad ,\theta,\varphi,t > 0. \tag{3.11}$$

The following lemma shows that the normalized form always exists.

Lemma 3.1 *The distribution function of any non-negative random variable with finite moments can always be expressed in a normalized form.*

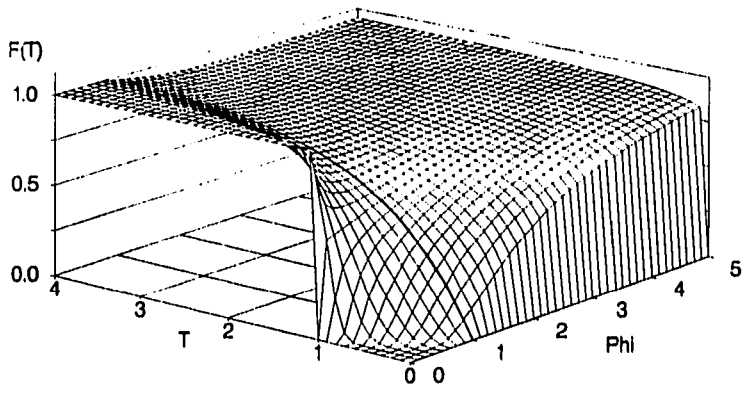
[2]Alternatively, one could argue the location parameter is fixed at zero.

Table 3.1: Conventional Descriptions of Four Non-negative Distribution Functions

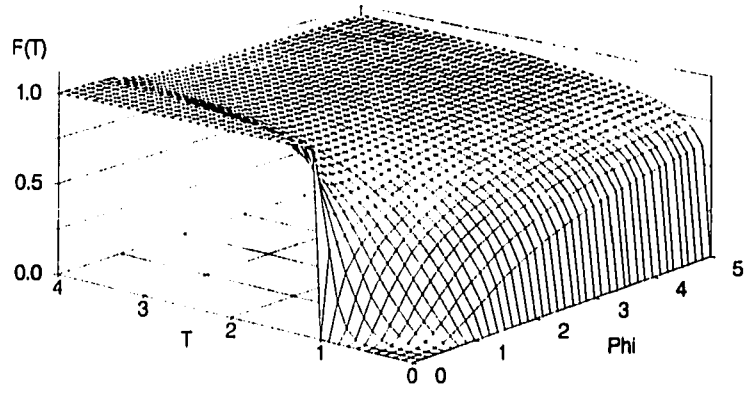
Distribution	Subscript	f(t)	Notes
Exponential	M	$\frac{1}{\theta}\exp\left[-\left(\frac{t}{\theta}\right)\right]$	$\theta > 0, \varphi = 1$
Gamma	Γ	$\frac{t^{\alpha-1}\exp\left(-\frac{t}{\beta}\right)}{\beta^{\alpha}\Gamma(\alpha)}$	$\alpha, \beta > 0$
Weibull	W	$\frac{\alpha t^{\alpha-1}\exp\left[-\left(\frac{t}{\beta}\right)^{\alpha}\right]}{\beta^{\alpha}}$	$\alpha, \beta > 0$
Lognormal	N	$\frac{1}{t\sqrt{2\pi\alpha}}\exp\left(-\frac{\ln t-\beta}{2\alpha}\right)$	$\alpha > 0, \beta$ unrestricted

Proof: Since T is a non-negative random variable with finite moments, one can define $\Phi = (\theta, \varphi_1, \varphi_2, \ldots)$, where $\theta = \mathrm{E}[T]$ and $\varphi_n = \frac{(\mathrm{E}[(T-\theta)^n])^{1/n}}{\theta}$. This vector contains all information about the distribution function of T. Clearly, all parameters except θ are unitless, so the lemma is proven. □

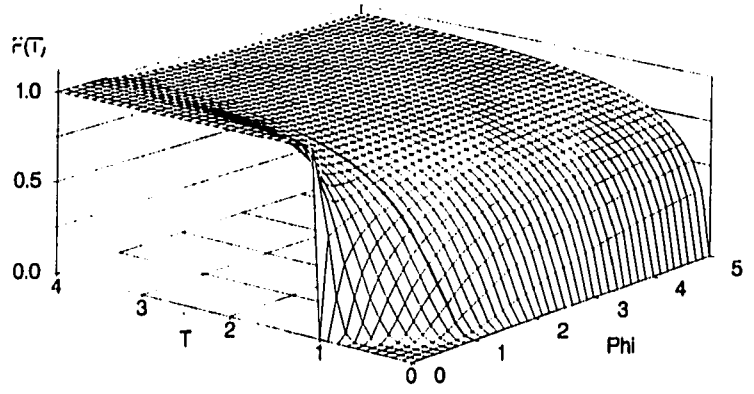
One can state any non-negative distribution in terms of $\Phi = (\theta, \varphi_1, \varphi_2, \ldots)$, where $\theta = \mathrm{E}[T]$ and φ_k; $k = 1, 2, \ldots$ are unitless shape parameters. Although Φ is an infinite-element vector, a n-parameter distribution function will be defined by the first n elements of Φ. The remaining elements of Φ will be redundant and may be discarded. Table 3.1 lists the usual definitions of several common distribution models [44] and Table 3.2 shows the relationships between the conventional and normalized parameters. Figure 3.1 illustrates three of these distributions as functions of θ and φ.



a. $F_{\Gamma}(t;\theta,\varphi)$



b. $F_{W}(t;\theta,\varphi)$



c. $F_{N}(t;\theta,\varphi)$

Figure 3.1: Some Probability Distributions in Normalized Form

Table 3.2: Relations Among Distribution Parameter Sets

Sub-Script	Conventional Parameters		Normalized Parameters	
	α	β	θ	φ
Γ	$\frac{1}{\varphi^2}$	$\theta\varphi^2$	$\alpha\beta$	$\sqrt{\frac{1}{\alpha}}$
W	$\frac{\Gamma\left(\frac{2}{\alpha}+1\right)}{\left[\Gamma\left(\frac{1}{\alpha}+1\right)\right]^2} = 1+\varphi^2$	$\frac{\theta}{\Gamma\left(\frac{1}{\alpha}+1\right)}$	$\beta\Gamma\left(\frac{1}{\alpha}+1\right)$	$\sqrt{\frac{\Gamma\left(\frac{2}{\alpha}+1\right)}{\Gamma\left(\frac{1}{\alpha}+1\right)}-1}$
N	$\ln\left(1+\varphi^2\right)$	$\ln\theta - \frac{\ln\left(1+\varphi^2\right)}{2}$	$\exp\left(\beta+\frac{\alpha}{2}\right)$	$\sqrt{\exp(\alpha)-1}$

Characteristics of general non-negative random variables

A major advantage to normalization is that one can more easily compare one distribution to another. In particular, the fact that θ is the only scale parameter leads to a simple, but useful result.

Theorem 3.1 *The distribution function of any non-negative random variable T with finite moments is a non-increasing function of the mean.*

Proof: By Lemma 3.1 one may express the distribution of T as $\mathrm{F}(t;\Phi)$ where $\Phi = (\theta, \varphi_1, \varphi_2, \ldots)$. Because the choice of time units is arbitrary, $\mathrm{F}(t;\Phi) = \mathrm{F}(x;\Phi_N)$ where $x = t/\theta$ and $\Phi_N = (1, \varphi_1, \varphi_2, \ldots)$. Because $\mathrm{F}(x)$ is a non-decreasing function of x, by the chain rule, $\mathrm{F}(t)$ must be a non-increasing function of θ. □

This argument also shows that plots in Figure 3.1 are actually plots of $\mathrm{F}(t/\theta;\varphi)$, and without loss of generality, one may assume $\theta = 1.0$.

Bounds on the distribution of T This next theorem establishes a lower bound on $\mathrm{F}(t)$.

Theorem 3.2 (Markov Lower Bound on $\mathrm{F}(t)$) *If T is a non-negative random variable with finite mean θ then,*

$$F(t;\theta,\varphi) \;\geq\; F_{\mathrm{MLB}}(t;\theta) \;\triangleq\; \begin{cases} 0 & ,\, 0 < t \leq \theta \\ 1 - \frac{\theta}{t} & ,\, t \geq \theta. \end{cases}$$

Proof: First, one need consider only $t > \theta$ since $\mathrm{F}(t) \geq 0$. Then, the result follows directly from Markov's inequality [45, p. 61]:

$$\Pr\{X > a\} \;\leq\; \frac{\mathrm{E}[X]}{a}$$

with $X = T$ and $a = t$. □

The next theorem presents upper and lower bounds based on Chebyshev's inequality. Unlike the Markov bounds, these hold for all random variables with finite moments.

Theorem 3.3 (Chebyshev's Bounds on $\mathrm{F}(x)$) *If x is any random variable with finite mean μ and finite variance σ^2, then:*

$$F(x;\mu,\sigma) \;\geq\; F_{\mathrm{CLB}}(x;\mu,\sigma) \;\triangleq\; \begin{cases} 0 & ,\, x \leq \mu + \sigma \\ 1 - \left(\frac{\sigma}{x-\mu}\right)^2 & ,\, x \geq \mu + \sigma \end{cases}$$

and

$$F(x;\mu,\sigma) \;\leq\; F_{\mathrm{CUB}}(x;\mu,\sigma) \;\triangleq\; \begin{cases} 1 & ,\, x \geq \mu - \sigma \\ \left(\frac{\sigma}{\mu-x}\right)^2 & ,\, x \leq \mu - \sigma. \end{cases}$$

Proof: Chebyshev's inequality [45] states that for any $k > 0$,

$$\Pr\{|X - \mu| \geq k\} \leq \left(\frac{\sigma}{k}\right)^2.$$

Let $k = x - \mu$. If $X > \mu$ then $k > 0$ and Chebyshev's inequality holds. Also, since $X - \mu$ is positive,

$$\Pr\{X - \mu \geq x - \mu\} \leq \left(\frac{\sigma}{x - \mu}\right)^2$$

which leads to the lower bound. On the other hand, if $k = \mu - x$ and $X < \mu$, then

$$\Pr\{\mu - X \leq \mu - x\} \leq \left(\frac{\sigma}{\mu - x}\right)^2$$

which is the upper bound. □

In the case of a non-negative random variable T with mean θ and coefficient of variation φ,

$$\mathrm{F_{CLB}}(t;\theta,\varphi) \triangleq \begin{cases} 0 & , 0 < t \leq \theta(1+\varphi) \\ 1 - \left(\frac{\theta\varphi}{t-\theta}\right)^2 & , t \geq \theta(1+\varphi) \end{cases}$$

and

$$\mathrm{F_{CUB}}(t;\theta,\varphi) \triangleq \begin{cases} 1 & , t \geq \theta(1-\varphi) \\ \left(\frac{\theta\varphi}{\theta-t}\right)^2 & , 0 < t \leq \theta(1-\varphi). \end{cases}$$

Figure 3.2 illustrates these bounds for $\varphi = 1$ together with the lognormal and exponential [3] distribution functions.

Characteristics of low-variance non-negative random variables

Figure 3.3 illustrates several distributions together with bounds on $\mathrm{F}(t)$ with $\varphi = 0.25$. In this range these distributions are very similar and the Chebyshev

[3] The Weibull and gamma distributions coincide with the exponential when $\varphi = 1$.

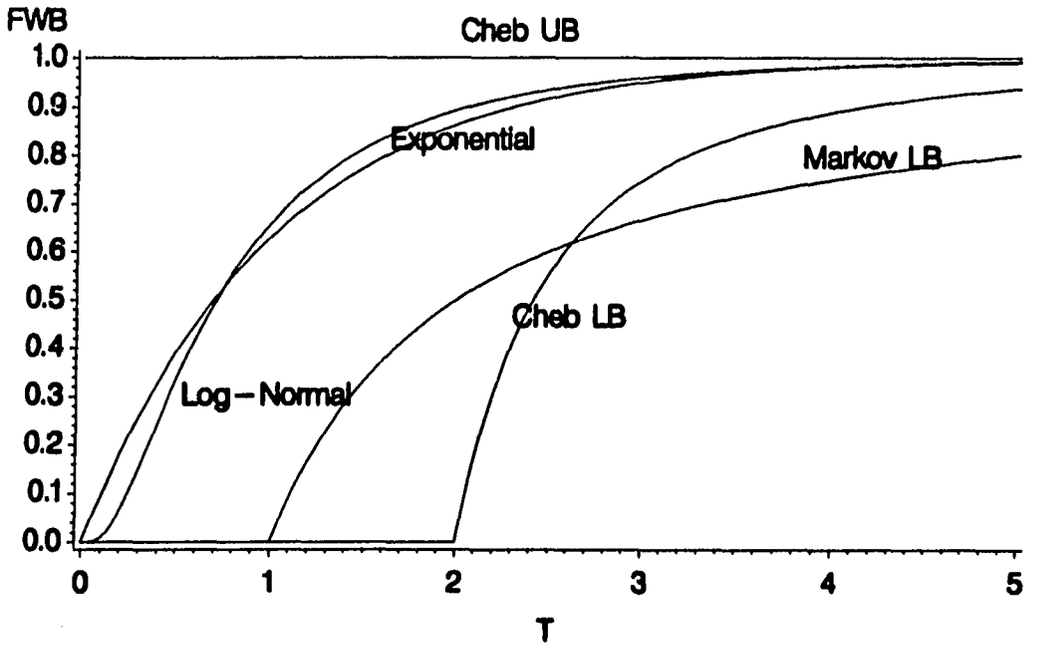


Figure 3.2: Bounds on $F(t;\theta,\varphi)$ for $\theta = 1$ and $\varphi = 1$

bounds dominate. Although the Chebyshev bounds seem loose, they do confirm that $F(t)$ converges to

$$F_d(t;\theta) = \begin{cases} 0 & ,t < \theta \\ 1 & ,t \geq \theta \end{cases} \tag{3.12}$$

as $\varphi \to 0$. In addition, they imply that if φ is small enough, $F(t)$ increases with φ for $t < \theta$ and decreases with t for $t > \theta$.

K_m distributions Suppose T is the sum of m independent, but not necessarily identical, exponential random variables with means $\theta_1, \theta_2, \ldots, \theta_m$. This distribution of T is often termed the *generalized Erlang* and will be denoted K_m below. It differs from the m-phase Erlang distribution in that the m independent random variables need not have the same mean.

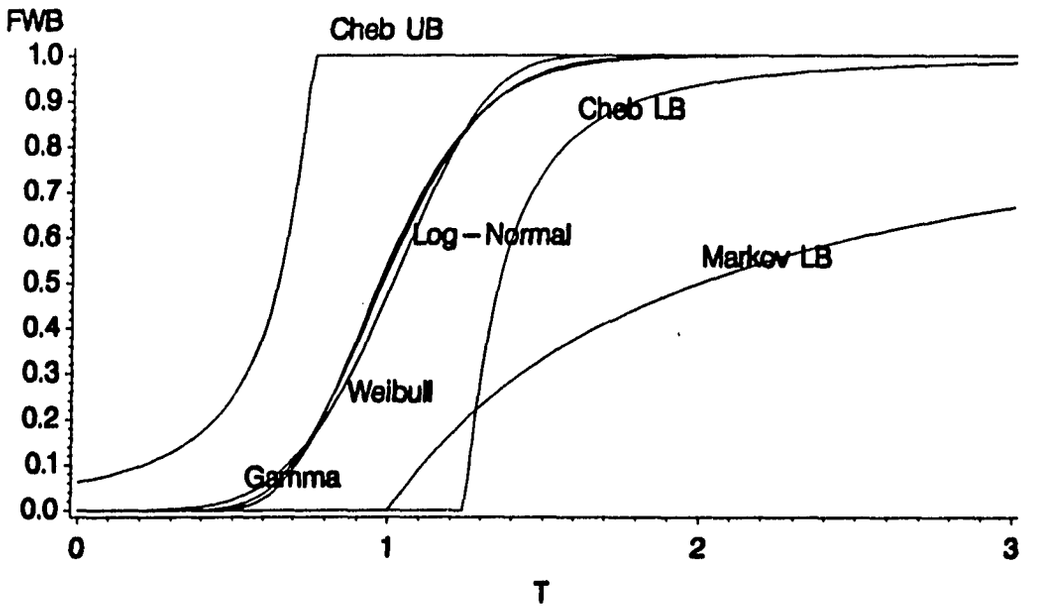


Figure 3.3: Bounds on F(t) when $\varphi = 0.25$

One may define K_m distributions in several ways. The following defines $\mathrm{F}_{\mathrm{K}m}(t)$ in terms of $\Theta = (\theta_1, \theta_2, \ldots, \theta_m)$, a vector of the individual phase mean times. Because the order of the index is arbitrary, assume $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_m$. Because T is the sum of m random variables, $\theta = \mathrm{E}[T] = \sum_{i=1}^{m} \theta_i$. Because the i-th random variable is exponential with variance θ_i^2 and the variables are independent, $\mathrm{Var}(T) = \sum_{i=1}^{m} \theta_i^2$. Also, for the same reasons, the Laplace- Stiltjes transform (LST) of the K_m distribution function is:

$$\mathrm{F}^*_{\mathrm{K}m}(s;\Theta) = \prod_{i=1}^{m} \left(\frac{1}{1+s\theta_i} \right). \tag{3.13}$$

One can determine $d\mathrm{F}_{\mathrm{K}m}(t)$ by partial-fraction expansion and term-by-term inversion of (3.13).

If the θ_i are all equal, $\mathrm{F}^*(s) = [m/(m+\theta s)]^m$ where $\theta = m\theta_i$. Hence T is a

m-phase Erlang random variable,

$$dF_{Km}(t;\theta) = \left(\frac{m}{\theta}\right)^m \frac{t^{m-1}}{(m-1)!} \exp\left(-\frac{mt}{\theta}\right) dt \tag{3.14}$$

and

$$F_{Km}(t;\Theta) = 1 - \exp\left(-\frac{mt}{\theta}\right) \sum_{i=0}^{m-1} \frac{1}{i!}\left(\frac{mt}{\theta}\right)^i. \tag{3.15}$$

On the other hand, if the θ_i are distinct, then

$$F^*_{Km}(s;\Theta) = \sum_{i=1}^{m} \left(\frac{A_i(\Theta)}{1+s\theta_i}\right) \tag{3.16}$$

and

$$dF_{Km}(t;\Theta) = \left(\frac{1}{\theta}\right) \sum_{i=1}^{m} A_i(\Theta)\left(\frac{1}{k_i}\right) \exp\left(-\frac{t}{k_i\theta}\right) dt \tag{3.17}$$

where[4] $k_i = \theta_i/\theta$, and

$$A_i(\Theta) = \prod_{\substack{n=1 \\ n\neq i}}^{m} \left(\frac{k_i}{k_i - k_n}\right). \tag{3.18}$$

Because the A_i arise from the expansion of a fraction with numerator 1.0, their sum is 1.0 and $\sum_{i=1}^{m} A_i/k_i = 0$. Since the A_i sum to one,

$$F_{Km}(t;\Theta) = 1 - \sum_{i=1}^{m} A_i(\Theta) \exp\left(-\frac{t}{\theta_i}\right). \tag{3.19}$$

Since $k_1 > \cdots > k_m$, A_1 is positive, the A_i alternate in sign, and $|A_1| > \cdots |A_m|$.

The following shows that (3.19) is the limit of (3.15) as the $\theta_i \to \theta/m$. When $m = 2$,

$$\frac{dF_{K2}(t;\theta_1,\theta_2)}{dt} = \left(\frac{k_1}{k_1-k_2}\right)\left(\frac{1}{\theta_1}\right)\exp\left(-\frac{t}{\theta_1}\right) + \left(\frac{k_2}{k_2-k_1}\right)\left(\frac{1}{\theta_2}\right)\exp\left(-\frac{t}{\theta_2}\right)$$

[4]This definition of A_i differs from that in [46, 47, 48] by the factor $1/k_i$.

$$= \frac{\exp\left(-\frac{t}{\theta_1}\right) - \exp\left(-\frac{t}{\theta_2}\right)}{\theta_1 - \theta_2} \tag{3.20}$$

Let $\delta = \theta_1 - \theta/2 = \theta/2 - \theta_2$. Then

$$\frac{dF_{K_2}(t;\theta_1,\theta_2)}{dt} = \frac{\exp\left(-\frac{t}{\theta/2+\delta}\right) - \exp\left(-\frac{t}{\theta/2-\delta}\right)}{2\delta}. \tag{3.21}$$

Applying L'Hôpital's rule for ratios,

$$\lim_{\delta\to 0} \frac{dF_{K_2}(t)}{dt} = \lim_{\delta\to 0} \frac{\frac{4t\exp\left(-\frac{t}{\theta/2+\delta}\right)}{(\theta+2\delta)^2} + \frac{4t\exp\left(-\frac{t}{\theta/2-\delta}\right)}{(\theta-2\delta)^2}}{2}.$$

$$= \left(\frac{2}{\theta}\right)^2 t\exp\left(-\frac{2t}{\theta}\right). \tag{3.22}$$

For larger values of m, one must use L'Hôpital's rule $m-1$ times which produces the $(m/\theta)^m\, t^{m-1}/(m-1)!$ term. Since this agrees with (3.14), expectations in this limiting case will equal the limits of the expectations [49]. Thus, without loss of generality, one can assume that the θ_i are distinct.

Because $\sum_{i=1}^{m} \frac{A_i(\Theta)}{k_i} = 0$, $f_{K_m}(0) = 0$. Because all $k_i \geq 0$, and $\sum_{i=1}^{r} k_i = 1$,

$$\varphi^2 = \sum_{i=1}^{m} k_i^2 \tag{3.23}$$

cannot be greater than one. Furthermore, (3.23) shows that φ reaches its maximum when $k_1 \to 1.0$, $k_i \to 0; i = 2,\ldots,m$ and its minimum when $k_i = 1/m; i = 1,\ldots,m$. Thus, if T is a K_m random variable, $\sqrt{1/m} \leq \varphi \leq 1$. Finally, the limit as $r \to \infty$ and $\varphi \to 0$ is the deterministic distribution $T \equiv \theta$.

Utility of K_m distributions Because one need not have equal phase means, K_m distributions are more general models of multi-phase processes than the Erlang. However, because φ is restricted to the range $1/\sqrt{m} \leq \varphi < 1$, the analyst will have to use a very large m if the variance of the distribution is very low. In spite of this practical limitation, it is a useful theoretical distribution, since given a sufficiently large m, one can closely approximate any low-variance distribution with a K_m distribution.

Characteristics of high-variance non-negative random variables

Bounds on F(t) Figure 3.4 shows the Markov bound, together with several distributions for $\varphi = 2$. Because $\mathrm{F_{MLB}}(\cdot)$ is independent of φ, it is a useful bound even if φ is very large. When $\varphi > 1$, the differences among the distributions seem to be much more pronounced for $t < \theta$. The B_r distribution family will be used to illustrate this point.

B_r distributions If T is a stochastic mix of r exponential random variables, one can view it as the result of a branching process and

$$\mathrm{f_{B_r}}\left(t; \Phi_{\mathrm{B}r}\right) = \sum_{j=1}^{r} p_j \left(\frac{1}{\theta_j}\right) \exp\left(-\frac{t}{\theta_j}\right) \tag{3.24}$$

where $\Phi_{\mathrm{B}r} = (\theta_1, \theta_2, \ldots, \theta_r, p_1, p_2, \ldots, p_r)$. Here θ_j is the mean of the exponential random variable associated with the j-th branch and p_j is the probability of selecting that branch [26, p. 36]. Because one may combine two branches if their means are equal, one may assume the θ_j are distinct. Of course, $\forall j, p_j \geq 0$ and $\theta_j > 0$. Also,

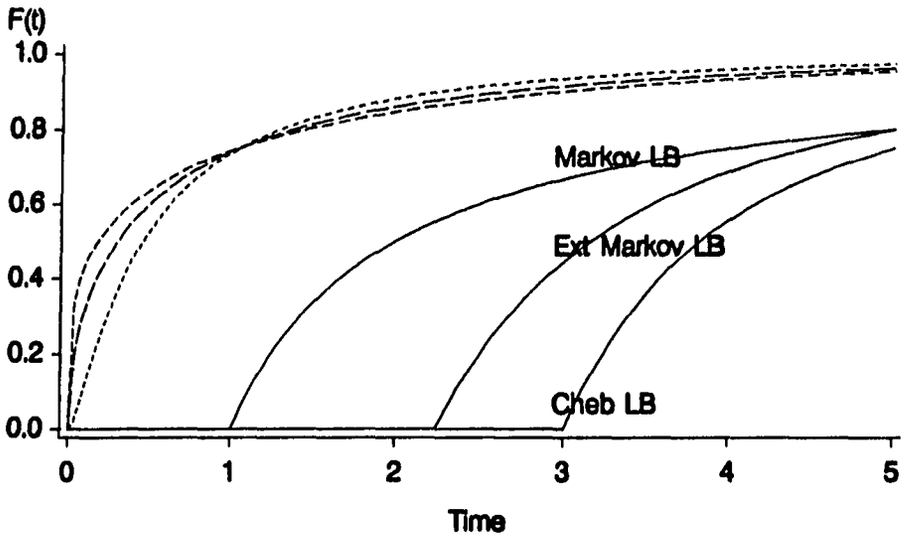


Figure 3.4: Bounds and Several Examples of High-Variance Distributions

$\sum_{j=1}^{r} p_j = 1$. Additionally,

$$\mathrm{F}_{\mathrm{B}r}\left(t; \Phi_{\mathrm{B}r}\right) = \sum_{j=1}^{r} p_j \left[1 - \exp\left(-\frac{t}{\theta_j}\right)\right] \tag{3.25}$$

is a convex combination of the r branch exponential distribution functions. Figure 3.5 shows an example of $\mathrm{F}_{\mathrm{B}r}(\cdot)$ together with the exponential distributions for each of its branches. In this case, $\theta_1 = 0.25$, $\theta_2 = 3$, $p = 8/11$, and $q = 3/11$. Overall, $\theta = 1$ and $\varphi = 2$. For comparison, this figure also shows $\mathrm{F}_{\Gamma}(t; 1, 2)$.

Properties of B_r Distributions Conditioning on the branch taken shows $\mathrm{E}[T] = \theta = \sum_{j=1}^{r} p_j \theta_j$. Let $w_j \stackrel{\Delta}{=} \theta_j/\theta$. Then $\theta = \sum_{j=1}^{r} p_j w_j \theta$ and $\sum_{j=1}^{r} p_j w_j = 1$. Since $p_r = 1 - \sum_{j=1}^{r-1} p_j$ and $w_r = \left(1 - \sum_{j=1}^{r-1} p_j w_j\right)/p_r$, one can define $\mathrm{F}_{\mathrm{B}r}(\cdot)$

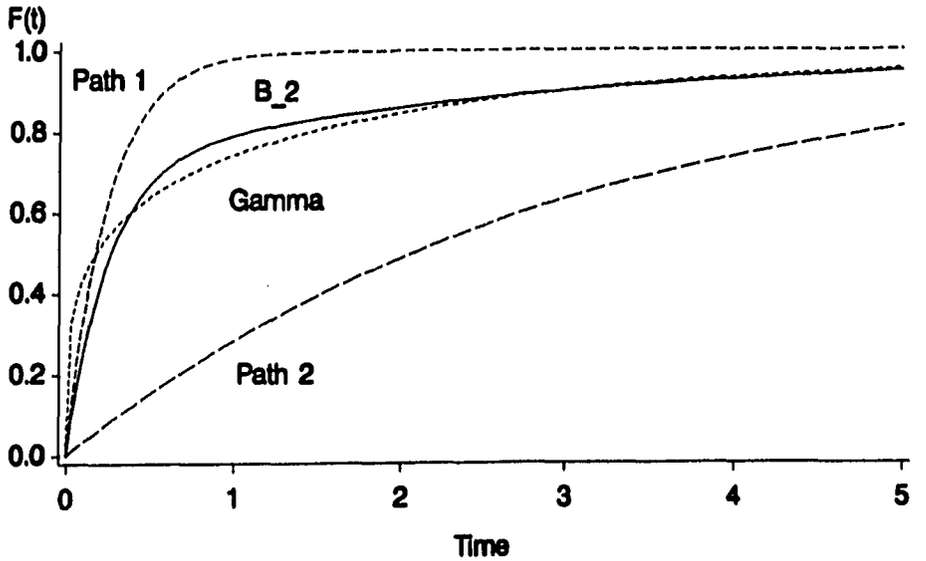


Figure 3.5: Components of a B_2 Distribution

using the $2r-1$ element parameter set

$$\Phi_{\mathrm{B}r} = \left(\theta, w_1, w_2, \ldots, w_{r-1}, p_1, p_2, \ldots, p_{r-1}\right). \tag{3.26}$$

Because elements of this set except θ are unitless, this is a normalized form. Among other things, this means that without loss of generality, one can assume $\theta = 1$. In addition, since $\mathrm{Var}[T] = \mathrm{E}[T^2] - \mathrm{E}[T]^2$,

$$\begin{aligned} \varphi &= \frac{\mathrm{E}\left[T^2\right]}{\theta^2} - 1 \\ &= \sum_{j=1}^{r} p_j \left(2w_j^2\right) - 1. \end{aligned} \tag{3.27}$$

Thus, one can express $\Phi_{\mathrm{B}r}$ in terms of θ, φ, and some other unitless parameters, such as $p_1, p_2, \ldots, p_{r-1}$ and $w_1, w_2, \ldots, w_{r-2}$. The following theorem proves that B_r distributions are necessarily high-variance distributions.

Theorem 3.4 *Given any positive* $\theta_1, \theta_2, \ldots, \theta_r$ *and any probabilities* $p_1, p_2, \ldots, p_r$ *such that* $p_j > 0; j = 1, 2, \ldots, r$ *and* $\sum_{j=1}^{r} p_j = 1$. *If*

$$F_{\mathrm{T}}(t; \Phi_n) = 1 - \sum_{j=1}^{r} p_j \exp\left(-\frac{t}{\theta_j}\right)$$

is the distribution of T, *then the coefficient of variation is greater than one.*

Proof: Without loss of generality, let $\theta = 1$, so $w_j = \theta_j$. Let $u_j = w_j - 1$. Then since $\sum_{j=1}^{r} p_j w_j = 1$,

$$\begin{aligned} 1 &= \sum_{j=1}^{r} p_j \left(1 + u_j\right) \\ &= \sum_{j=1}^{r} p_j + \sum_{j=1}^{r} p_j u_j \end{aligned} \tag{3.28}$$

which shows $\sum_{j=1}^{r} p_j u_j = 0$. Thus

$$\begin{aligned} \sum_{j=1}^{r} p_j w_j^2 &= \sum_{j=1}^{r} p_j \left(1 + u_j\right)^2 \\ &= 1 + \sum_{j=1}^{r} p_j u_j^2 \end{aligned} \tag{3.29}$$

and substituting (3.29) into (3.27) shows

$$\varphi^2 = 1 + 2 \sum_{j=1}^{r} p_j u_j^2 \tag{3.30}$$

which cannot be less than one. □

A normalized B_2 distribution Let $\theta = 1$ and $r = 2$. Then

$$\mathrm{F}_{\mathrm{B}_2}(t; 1, p, w) = 1 - p \exp\left(-\frac{t}{w}\right) - (1 - p) \exp\left(-\frac{t}{w_2}\right) \tag{3.31}$$

with $w_2 = (1 - wp)/(1 - p)$. From (3.27),

$$\varphi^2 = 2\left[pw^2 + (1-p)\left(\frac{1-pw}{1-p}\right)^2\right] - 1. \tag{3.32}$$

Thus

$$p(w,\varphi) = \frac{\varphi^2 - 1}{2w^2 - 4w + \varphi^2 + 1}. \tag{3.33}$$

Now, since both weights must be positive, $1 - wp > 0$. This implies that $p < 1/w$. Substituting $1/w$ for p into (3.33) yields

$$2w^2 - \left(3 + \varphi^2\right) w + \varphi^2 + 1 > 0. \tag{3.34}$$

This convex quadratic has roots at 1 and $\left(\varphi^2 + 1\right)/2$. Since $\left(\varphi^2 + 1\right)/2 = 1 + \left(\varphi^2 - 1\right)/2$, the second root is greater than the first for $\varphi > 1$. Thus, one may have either $0 < w < 1$ or $w > \left(\varphi^2 + 1\right)/2$. Repeating the same argument with w_2 and $1 - p$ in place of w and p shows that either $0 < w_2 < 1$ or $w_2 > \left(\varphi^2 + 1\right)/2$. If $w < 1$ then

$$w_2 = \frac{1 - wp}{1 - p} > 1. \tag{3.35}$$

Conversely, if $w > 1$ then $w_2 < 1$. Thus, one branch mean will be in the interval $(0, 1)$ and the other in $(\frac{\varphi^2+1}{2}, +\infty)$. Because the numbering of the branches is arbitrary, assume $w \in (0, 1)$. Thus, for any $\varphi > 1$ one can construct a B_2 distribution by:

1. Selecting $w \in (0, 1)$,
2. Computing p from (3.33),
3. Computing w_2 from (3.35), and
4. Calculating probabilities from (3.31).

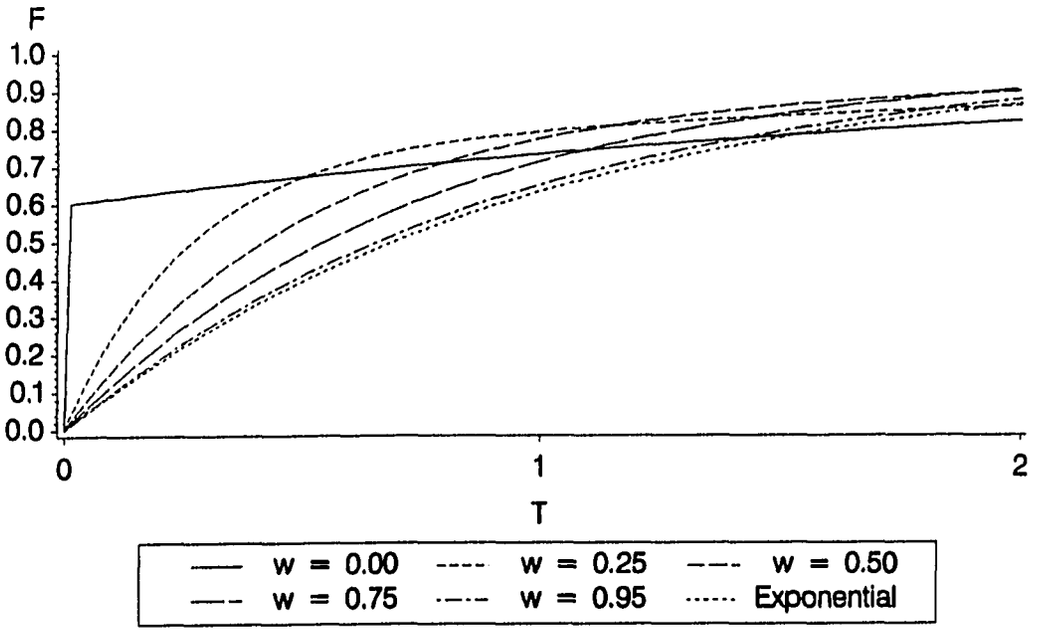


Figure 3.6: Examples of B_2 Distributions with $\varphi = 2$

Diversity of B_2 distributions For any $\varphi > 1$, there are an infinite number of B_2 distributions. Figure 3.6 illustrates the range of possible distributions when $\theta = 1$ and $\varphi = 2$. For comparison, this figure also shows the exponential (with $\varphi = 1$). It is clear that w has a great effect on $\mathrm{F}_{\mathrm{B}_2}(t;\theta,\varphi,w)$ and that as $w \to 1$, $\mathrm{F}_{\mathrm{B}_2}(t;\theta,\varphi,w)$ comes close to $\mathrm{F}_{\mathrm{M}}(t;\theta)$.

Summary

One may express the distribution of any non-negative random variable with finite moments in normalized form. In addition, by choice of scale, one may set $\theta = 1$ without changing any other normalized parameter value and the normalized distribution is a decreasing function of θ. One may bound $\mathrm{F}(t)$ using Markov and

Chebyshev bounds. Chebyshev bounds are most effective when $\varphi < 1$ and show that as $\varphi \rightarrow 0$, $F(t;\Phi) \longrightarrow F_d(t;\theta)$, the deterministic distribution. The Markov bound is most useful when $\varphi >> 1$. When φ is small, there is little difference in the shape of distribution functions with the same θ and φ. However, if φ is large, the variation is great. In particular, for any finite $\varphi > 1$, one may construct a distribution that is arbitrarily close to the exponential, but with heavier tails.

The Nature of Mean Timeliness

Introduction

This section explores the nature of mean timeliness and establishes some boundaries on its values. The analyst can use this information to bound the effectiveness of a system, or given ranges of acceptable performance, bound the range of significant performance values.

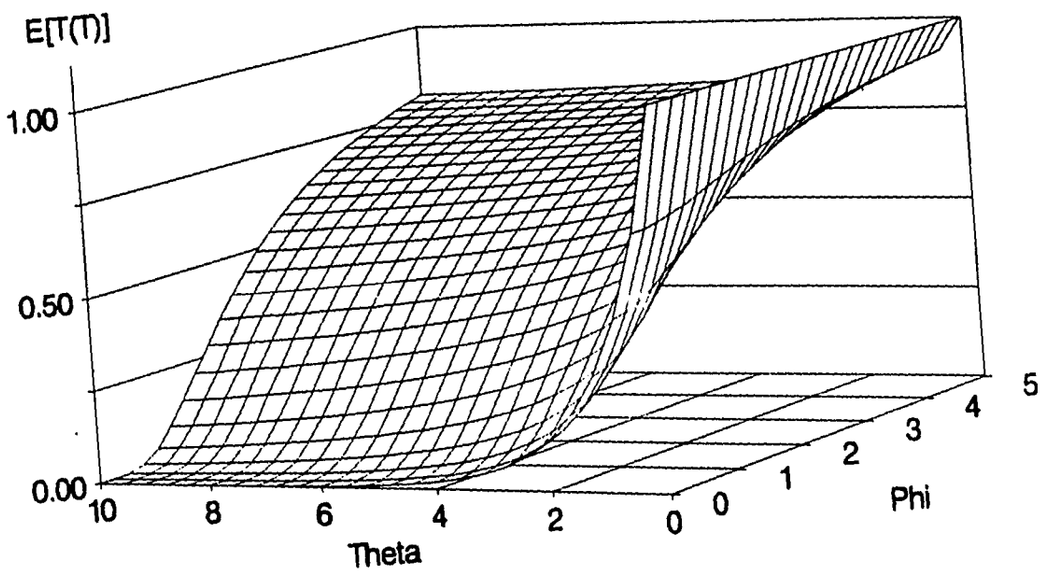
General features of mean timeliness

Figure 3.7 illustrates $E_\Gamma [\mathcal{T}_2 (T; 1, t_0) | \theta, \varphi]$. The obvious question is "What features of Figure 3.7 illustrate general features of $E[\mathcal{T}(T)]$ and what features are unique to $E_\Gamma \left[\mathcal{T}\left(T; \tilde{t}, t_0\right) \middle| \theta, \varphi\right]$?" This section formally establishes features of $E[\mathcal{T}(T)]$ for general timeliness functions and distributions as well as some specific results for the $\mathcal{T}_2(t; \tilde{t}, t_0)$ model. Table 3.3 summarizes these results.

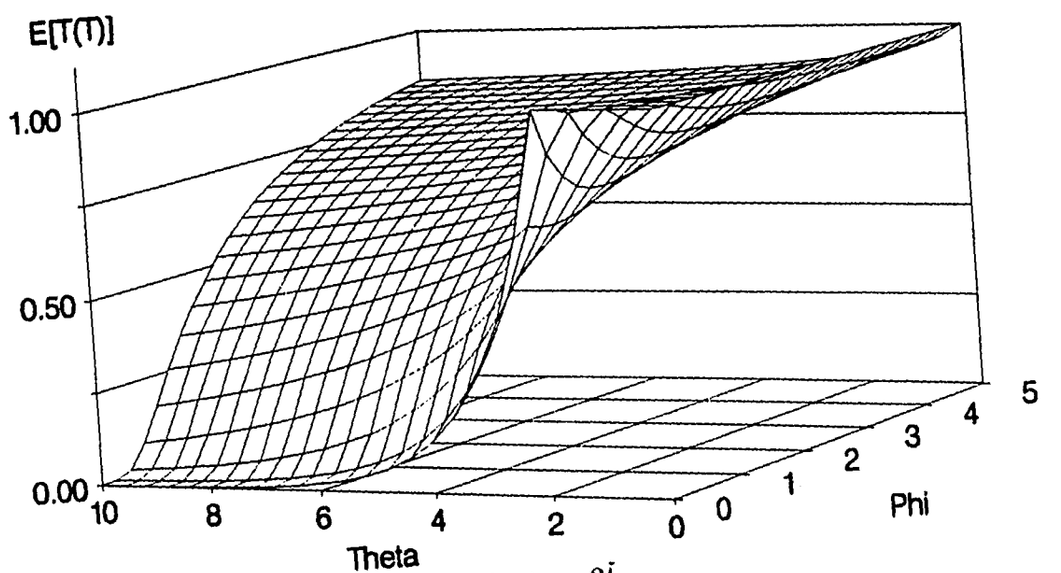
Mean timeliness is simply the expectation of $\mathcal{T}(t)$:

$$E_X [\mathcal{T}_Y (T; \Upsilon) | \Phi] = \int_0^\infty \mathcal{T}_Y (t; \Upsilon) \, dF_X (t; \Phi) . \tag{3.36}$$

Since $0 \leq \mathcal{T}(t) \leq 1$ and $\int_0^\infty dF(t) = 1$, the integral in (3.36) will converge. In addi-



a. $t_0 = 0$



b. $t_0 = 2i$

Figure 3.7: An Example of Mean Timeliness: $\mathrm{E}_{\Gamma}[T\ (T; 1, t_0) | \theta, \varphi]$

Table 3.3: Invariant Features of Mean Timeliness

$\forall t > 0, \mathcal{T}_a(t) \geq \mathcal{T}_b(t)$	$\Rightarrow$	$\mathrm{E}[\mathcal{T}_a(T)] \geq \mathrm{E}\left[\mathcal{T}_b(T)\right]$
$\forall t > 0, \mathrm{F}_1(t) \geq \mathrm{F}_2$	$\Rightarrow$	$\mathrm{E}_1[\mathcal{T}(T)] \geq \mathrm{E}_2[\mathcal{T}(T)]$
$\partial \mathrm{E}[\mathcal{T}(T)]/\partial\theta$	$\leq$	0
$\partial \mathrm{E}[\mathcal{T}_2(T)]/\partial\tilde{t}$	$\geq$	0
$\partial \mathrm{E}[\mathcal{T}_2(T)]/\partial t_0$	$\geq$	0

tion, both factors under the integral are positive and have independent parameters. This leads to a simple, but useful theorem.

Theorem 3.5 *For any non-negative random variable T, $\mathcal{T}_a(t) \geq \mathcal{T}_b(t)$ for all T implies that $E[\mathcal{T}_a(T)] \geq E[\mathcal{T}_b(T)]$.*

Proof:

$$\begin{aligned} \mathrm{E}[\mathcal{T}_a(T;\Upsilon)\,|\Phi] &= \int_0^\infty \mathcal{T}_a(t;\Upsilon)\, d\,\mathrm{F}_X(t;\Phi) \\ &\geq \int_0^\infty \mathcal{T}_b(t;\Upsilon)\, d\,\mathrm{F}_X(t;\Phi) \\ &= \mathrm{E}\left[\mathcal{T}_b(T;\Upsilon)\,|\Phi\right]. \end{aligned}$$

□

For example, because $\mathcal{T}_2(t)$ is an increasing function of both $\tilde{t}$ and t_0, $\mathrm{E}[\mathcal{T}_2(T)]$ must be, too. In a similar vein, one can show the effect of a dominant distribution.

Theorem 3.6 *For any non-negative random variable T, $F_1(t) \geq F_2(t)$ for all T implies that $E_1[\mathcal{T}(T)] \geq E_2[\mathcal{T}(T)]$.*

Proof: Integrating (3.36) by parts yields:

$$\mathrm{E}_{\mathrm{X}}\left[\mathcal{T}_{\mathrm{Y}}\left(T;\Upsilon\right)|\Phi\right] \;=\; -\int_0^{\infty} \mathrm{F}_{\mathrm{X}}\left(t;\Phi\right) d\mathcal{T}_{\mathrm{Y}}\left(t;\Upsilon\right).$$

Bearing in mind that $d\mathcal{T}_{\mathrm{Y}}(\cdot)$ is negative, the argument of Theorem 3.5 will also prove this theorem. □

The following theorem builds upon Theorem 3.6 and capitalizes on the special nature of normalized distributions.

Theorem 3.7 *For any non-negative random variable T with finite moments and any non-increasing timeliness function, $\mathcal{T}(t;\Upsilon)$, the mean timeliness is a decreasing function of $\theta = E[T]$.*

Proof: Represent $\mathrm{F}_{\mathrm{T}}(t;\Phi)$ in normalized form with $\Phi = (\theta, \varphi_1, \varphi_2, \varphi_3, \ldots)$. Since θ is the only scale parameter, $\mathrm{F}(t;\Phi) = \mathrm{F}(t/\theta;\Phi_{\mathrm{N}})$ where $\Phi_{\mathrm{N}} = (1, \varphi_1, \varphi_2, \varphi_3, \ldots)$. From Theorem 3.1 $\mathrm{F}(t)$ is a decreasing function of θ, so by Theorem 3.6, $\mathrm{E}[\mathcal{T}(T)]$ must also be a decreasing function of θ. □

These theorems confirm much of the behavior shown in Figure 3.7. For one, $\mathrm{E}[\mathcal{T}(T)]$ is a decreasing function of θ. Secondly, the normalized form allows one to select a scale in which any one of θ, $\tilde{t}$, or t_0 are equal to one. (In the following examples, $\tilde{t}$ will be 1.0, unless otherwise stated).

General bounds on mean timeliness

This section presents several bounds on mean timeliness. The analyst can use these to bound system effectiveness as well as working ranges of θ and φ.

The Markov lower bound Although the Markov lower bound is not very tight, it is useful in that one need not estimate φ.

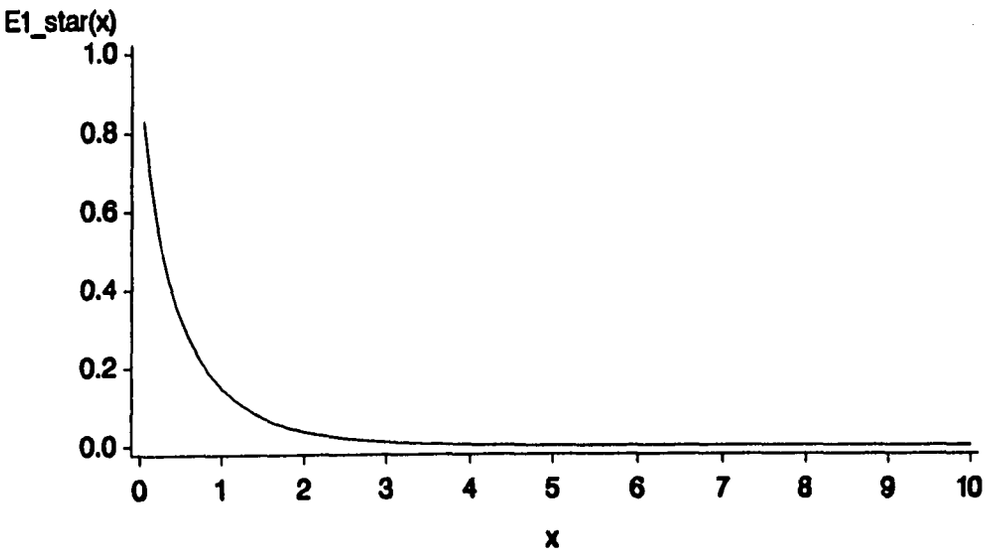


Figure 3.8: The Function $G_1(x) = \exp(-x) - xE_1(x)$

Corollary 3.1 (Markov Lower Bound on E[$\mathcal{T}(T)$]) *For any non-negative random variable T with finite mean θ and any non-increasing timeliness model $\mathcal{T}(t)$,*

$$\begin{aligned} E[\mathcal{T}(t)|\,\Phi] &\geq E_{\mathrm{MLB}}\,[\mathcal{T}(t)|\,\theta] \\ &= \mathcal{T}(\theta) + \theta \int_{\theta}^{\infty} \frac{d\mathcal{T}(t)}{t}. \end{aligned}$$

Proof: From Theorems 3.2 and 3.6,

$$\begin{aligned} \mathrm{E}\,[\mathcal{T}(t)|\,\Phi] &\geq \mathrm{E}_{\mathrm{MLB}}\,[\mathcal{T}(t)|\,\theta] \\ &= -\int_{\theta}^{\infty} \left(1 - \frac{\theta}{t}\right) d\mathcal{T}(t). \end{aligned}$$

Separating the terms within the parentheses and evaluating the first leads to the result. □

Theorem 3.1 establishes a characteristic lower bound for any given timeliness function. For example, since $d\mathcal{T}_2(t;1,t_0) = 0$ for $t \leq t_0$,

$$\begin{aligned} \mathrm{E}_{\mathbf{MLB}}\left[\mathcal{T}_2(T;1,t_0)|\,\theta\right] &= \mathcal{T}_2(\theta;1,t_0) - \theta \int_{\max(\theta,t_0)}^{\infty} \left(\frac{1}{t}\right) \exp\left[-(t-t_0)\right] dt \\ &= \begin{cases} 1 - \theta \exp(t_0)\, E_1(t_0) & , \theta \leq t_0 \\ \exp(t_0)\left[\exp(-\theta) - \theta E_1(\theta)\right] & , \theta > t_0 \end{cases} \end{aligned} \tag{3.37}$$

where

$$E_\nu(x) = \int_x^{+\infty} \exp(-u) u^{-\nu} du \tag{3.38}$$

is the *exponential integral.* Pagurova discusses this family of functions and presents extensive tables of their values in [50]. Figure 3.8 shows a plot of $G_1(\theta) = \exp(-\theta) - \theta E_1(\theta)$. Also, Figure 3.9 shows $\mathrm{E}_{\mathbf{MLB}}[\cdot]$.

Let $\theta_v^\star$ be the smallest value of θ for which the Markov lower bound is less than or equal to a given value v. If $\theta_v^\star \leq t_0$, (3.37) implies,

$$\theta_v^\star = \tilde{t}(1-v) C_{\mathbf{MLB}} \tag{3.39}$$

where

$$C_{\mathbf{MLB}} = \frac{\exp\left(-t_0/\tilde{t}\right)}{E_1\left(t_0/\tilde{t}\right)} \tag{3.40}$$

is a unitless quantity, depending only on the ratio of t_0 to $\tilde{t}$. Table 3.4 lists $C_{\mathbf{MLB}}$ as a function of $\mathcal{P}$ for the CDB data as well as $\mathrm{E}_{\mathbf{MLB}}{}^\star(\mathcal{P})$, the smallest value of $\mathrm{E}_{\mathbf{MLB}}\left[\mathcal{T}_2(T)|\,\theta\right]$ for which (3.39) holds. Since it is unlikely that $V_{\max}$ will be less than 0.75, (3.39) is useful for all $\mathcal{P}$ codes except "B". Because $\mathrm{E}_{\mathbf{MLB}}\left[\mathcal{T}_2(T;1,0)|\,\theta\right] =$

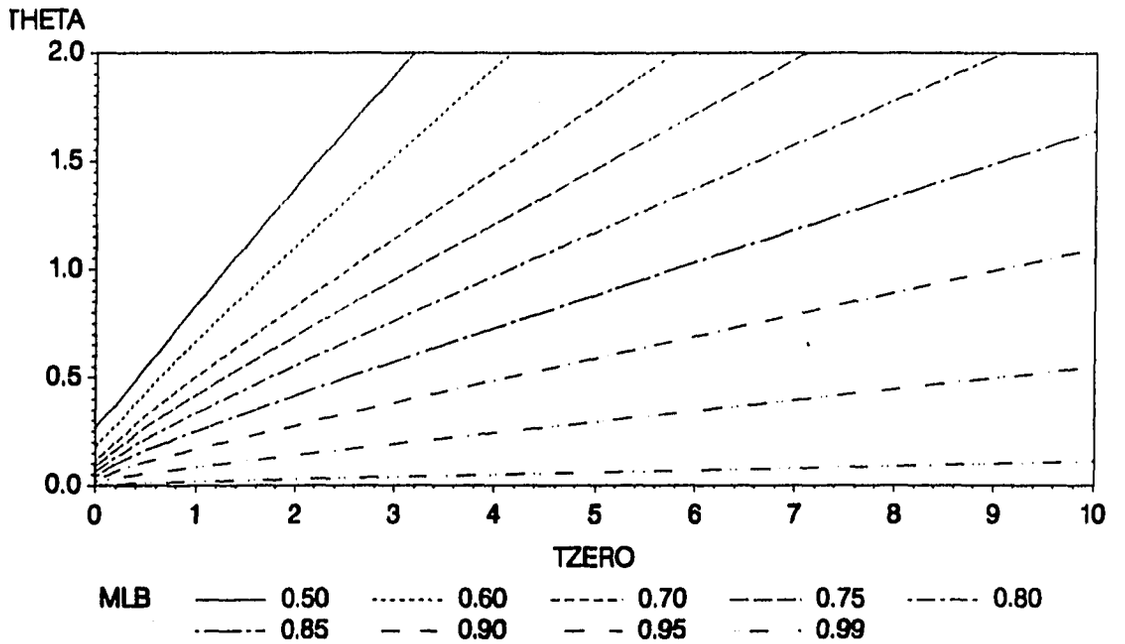


Figure 3.9: $\mathrm{E}_{\mathrm{MLB}}[\mathcal{T}_2(T)]$

$G_1(\theta)$, one could find $\theta_v^\star$ by iteration. However, since $\mathrm{E}_{\mathrm{MLB}}[\cdot] \approx 0.85$ for $\theta/\tilde{t} = 0.04$, or $\theta = 4.34$ msec, it is unlikely that a lower bound would be practical for $\mathcal{P} = B$.

Analysts can use $\theta_v^\star$ to roughly analyze data. For example, if $\mathcal{P} = 6$ and the analysts would not be able to differentiate among systems with $\mathrm{E}[\mathcal{T}_2(T)] > 0.95$, then they would accept any system for which $\theta < 4.096(1 - 0.95)0.5682$ minutes, about 7 seconds.

Chebyshev Upper Bound If $\varphi < 1$, one may also bound $\mathrm{E}[\mathcal{T}(T)]$ from above.

Corollary 3.2 (Chebyshev Upper Bound on $\mathrm{E}[\mathcal{T}(T)]$) *If T is a non-negative random variable with finite mean θ and coefficient of variation $\kappa \leq \varphi$, and $\mathcal{T}(t)$ is a*

Table 3.4: The Ratio of Critical Mean Time to Value for the Markov Lower Bound to $\mathrm{E}\left[\mathcal{T}_2\left(T;\tilde{t},t_0\right)\middle|\theta\right]$

$\mathcal{P}$	$C_{\mathrm{MLB}}(\mathcal{P})$	$\mathrm{E}_{\mathrm{MLB}}{}^{\star}(\mathcal{P})$	$\mathcal{P}$	$C_{\mathrm{MLB}}(\mathcal{P})$	$\mathrm{E}_{\mathrm{MLB}}{}^{\star}(\mathcal{P})$
0	1.6692	0.4049	6	0.5682	0.7558
1	2.8635	0.2695	7	2.1917	0.3320
2	7.3715	0.1201	8	2.3619	0.3137
3	7.8579	0.4542	9	2.4734	0.3023
4	2.8635	0.2695	A	1.0118	0.5612
5	0.8655	0.6140	B	0.0000	—

non-increasing timeliness function, then

$$\begin{aligned} E[\mathcal{T}(t)|\,\Phi] &\leq E_{\mathrm{CUB}}\left[\mathcal{T}(t)|\,\theta,\varphi\right] \\ &= \mathcal{T}\left[\theta\left(1-\varphi\right)\right] - (\theta\varphi)^2 \int_0^{\theta(1-\varphi)} \frac{d\mathcal{T}(t)}{(\theta-t)^2} \end{aligned}$$

Proof: First, by Theorem 3.3 and the converse of Theorem 3.6,

$$\mathrm{E}\left[\mathcal{T}(t)|\,\Phi\right] \leq -\int_0^{\theta(1-\varphi)} \left(\frac{\theta\varphi}{\theta-t}\right)^2 d\mathcal{T}(t) - \int_{\theta(1-\varphi)}^{\infty} d\mathcal{T}(t)$$

which shows the inequality is true for $\kappa = \varphi$. Next,

$$\frac{\partial \mathrm{E}_{\mathrm{CUB}}\left[\mathcal{T}(t)|\,\theta,\varphi\right]}{\partial\varphi} = \frac{\partial\mathcal{T}\left[\theta(1-\varphi)\right]}{\partial\varphi} - 2\theta^2\varphi\int_0^{\theta(1-\varphi)} \frac{d\mathcal{T}(t)}{(\theta-t)^2} + \theta^2\mathcal{T}\left[\theta(1-\varphi)\right]$$

which is positive, consisting of three positive terms. □

In the case of the $\mathcal{T}_2(t)$ model, when $t_0 < \theta(1-\varphi)$,

$$\begin{aligned} &\mathrm{E}_{\mathrm{CUB}}\left[\mathcal{T}_2\left(t;1,t_0\right)|\,\theta,\varphi\right] \\ &= \exp\left[-\theta(1-\varphi)\right] + \\ &\quad \theta^2\varphi^2\exp\left[-\left(\theta-t_0\right)\right]\left[\frac{\exp\left(\theta-t_0\right)}{\theta-t_0} - \frac{\exp(\theta\varphi)}{\theta\varphi} - \int_{\theta\varphi}^{\theta-t_0}\frac{\exp(x)}{x}dx\right]. \quad (3.41) \end{aligned}$$

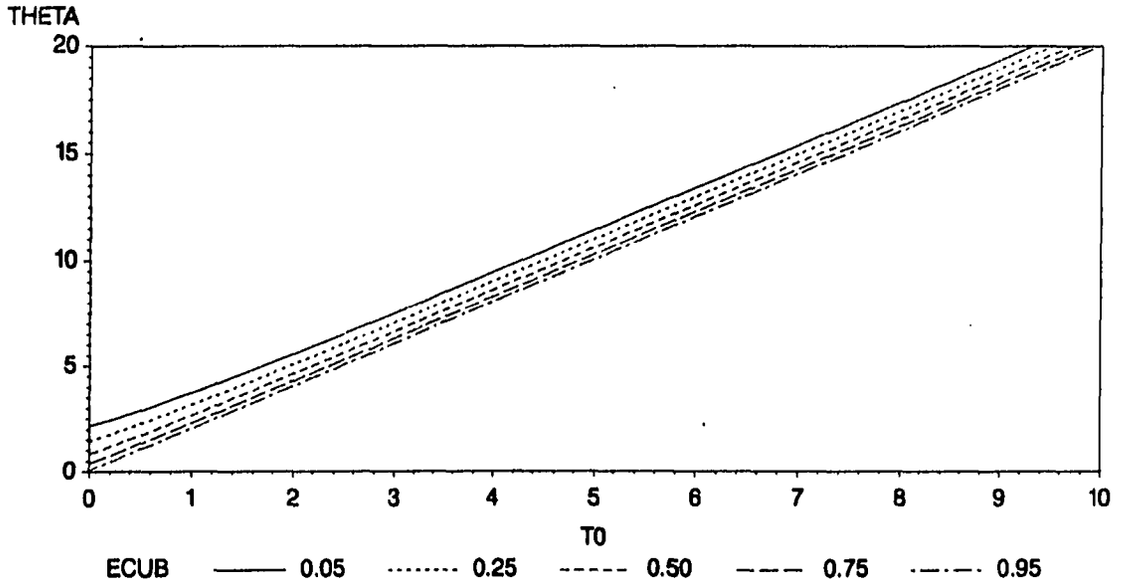


Figure 3.10: The Chebyshev Upper Bound on $\mathrm{E}[\mathcal{T}_2(T)]$ when $\varphi = 0.5$

Otherwise, $\mathrm{E}_{\mathrm{CUB}}[\cdot] = 1$. Thus, to use this bound, φ must be less than one and $t_0 < \theta(1-\varphi)$. One must also evaluate $\int_a^b \exp(x)/x\,dx$ numerically. This is not a serious problem unless φ is very small. Figure 3.10 illustrates $\mathrm{E}_{\mathrm{CUB}}[\mathcal{T}_2(T)]$ with $\varphi = 0.5$ evaluated using Simpson's rule with four interpolation points. The scale of θ is ten times greater than that in the earlier figures.

Mean timeliness when variance is low

When $\varphi < 1$, one may bound $\mathrm{E}[\mathcal{T}(T)]$ from above and below. One may also bound the working range of θ, given $\mathrm{E}_{\mathrm{min}}$ and $\mathrm{E}_{\mathrm{max}}$. Figure 3.11 shows upper bounds on θ and lower bounds on φ for integer values of t_0 assuming $\mathrm{E}_{\mathrm{min}} = 0.50$. If (θ, φ) lies above a line, mean timeliness will be below 0.5 when t_0 is the stated value. This figure is based on $\mathrm{E}_{\mathrm{CUB}}[\mathcal{T}_2(T)]$.

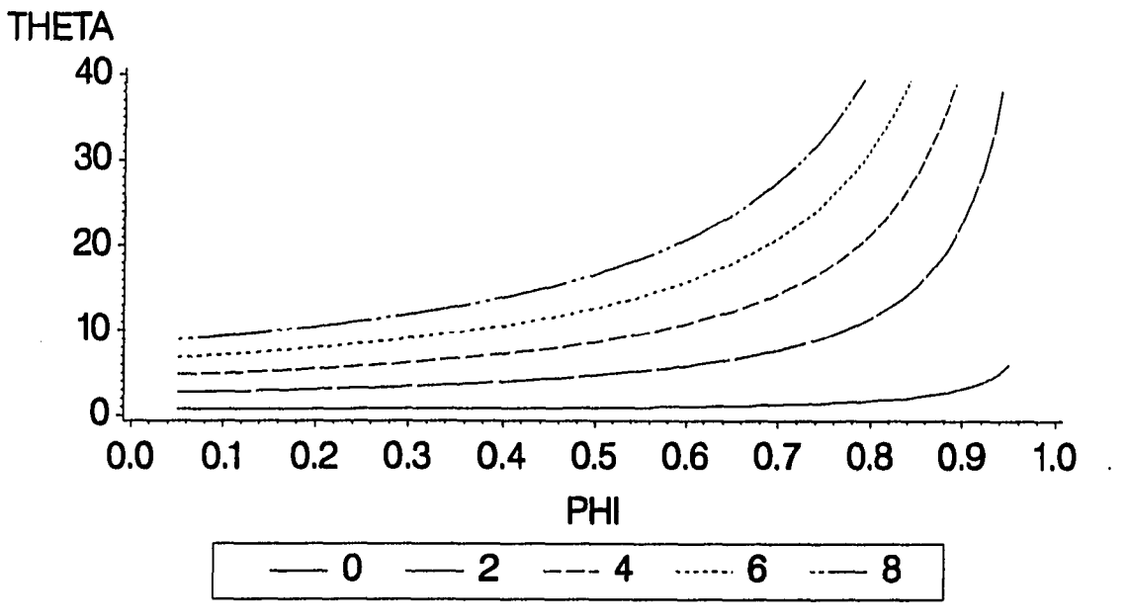


Figure 3.11: Chebyshev Upper Bounds on θ and Lower Bounds on φ for $\mathrm{E}[\mathcal{T}_2(T)]$ with $\mathrm{E}_{\min} = 0.50$

Mean timeliness for the K_m distribution If the θ_i are equal, $F_{\mathrm{K}m}(t)$ is a m-phase Erlang distribution, so

$$\mathrm{E}_{\mathrm{K}m}[\mathcal{T}(T)|\Theta] = \int_0^\infty \mathcal{T}(t)\left(\frac{m}{\theta}\right)^m \left(\frac{t^{m-1}}{(m-1)!}\right) \exp\left(-\frac{mt}{\theta}\right) dt. \quad (3.42)$$

However, if the θ_i are distinct,

$$\begin{aligned}\mathrm{E}_{\mathrm{K}m}[\mathcal{T}(T)|\Theta] &= \int_0^\infty \mathcal{T}(t) \sum_{i=1}^{m} A_i(\Theta)\left(\frac{1}{\theta_i}\right) \exp\left(-\frac{t}{\theta_i}\right) dt \\ &= \sum_{i=1}^{m} A_i(\Theta)\mathrm{E}_{\mathrm{M}}\left[\mathcal{T}(T)|\theta_i\right]. \end{aligned} \quad (3.43)$$

Because the θ_i are in order of decreasing size and $\mathrm{E}_{\mathrm{M}}[\mathcal{T}]$ is a decreasing function of θ, $\mathrm{E}_{\mathrm{M}}[\mathcal{T}(T)|\theta_1] \leq \mathrm{E}_{\mathrm{M}}[\mathcal{T}(T)|\theta_2] \leq \cdots \leq \mathrm{E}_{\mathrm{M}}[\mathcal{T}(T)|\theta_m]$. However, the A_i alternate in sign and decrease in absolute value with i. Thus, $\mathrm{E}_{\mathrm{K}m}[\mathcal{T}(T)]$ may either increase or decrease with increasing θ.

The results are elegantly simple for the $\mathcal{T}_2(t)$ model when $t_0 = 0$. Let $t = \sum_{i=1}^{m} t_i$, where $t_1, \ldots, t_m$ are the individual phase times. Then,

$$\begin{aligned} \mathcal{T}_2(t_1 + t_2 + \cdots + t_m; \tilde{t}, 0) &= \exp\left[-\sum_{i=1}^{m}\left(\frac{t_i}{\tilde{t}}\right)\right] \\ &= \prod_{i=1}^{m} \exp\left(-\frac{t_i}{\tilde{t}}\right). \end{aligned} \tag{3.44}$$

Since the T_i are independent,

$$\begin{aligned} \mathrm{E}_{\mathrm{K}m}\left[\mathcal{T}_2(T; \tilde{t}, 0)\right] &= \prod_{i=1}^{m} \mathrm{E}_{\mathrm{K}m} \exp\left(-\frac{T_i}{\tilde{t}}\right) \\ &= \prod_{i=1}^{m} \left(\frac{\tilde{t}}{\tilde{t} + \theta k_i}\right). \end{aligned} \tag{3.45}$$

This highlights the fact that in this particular case, the timeliness at the completion of Stage n depends only on $\mathcal{T}\left(\sum_{i=1}^{n-1} t_i\right)$, the timeliness at the beginning of the task, and t_n. This product will be greatest when $k_1 = 1, k_i = 0; i = 2, \ldots, m$ and smallest when all $k_i = 1/m$. Thus,

$$\left(\frac{\tilde{t}}{\tilde{t} + \frac{\theta}{m}}\right)^m \leq \mathrm{E}_{\mathrm{K}m}\left[\mathcal{T}_2(T; \tilde{t}, 0) \,|\Theta\right] \leq \frac{\tilde{t}}{\tilde{t} + \theta}. \tag{3.46}$$

Assuming the θ_i are distinct, (3.1) on p. 78 and (3.43) lead to:

$$\mathrm{E}_{\mathrm{K}m}\left[\mathcal{T}_2\left(T; \tilde{t}, t_0\right)\middle|\Theta\right] = 1 - \sum_{i=1}^{m} A_i(\Theta)\left(\frac{\theta_i}{\theta_i + \tilde{t}}\right) \exp\left(-\frac{t_0}{\theta_i}\right). \tag{3.47}$$

Since the A_i alternate in sign and can be very large, the analyst should not be surprised if $\mathrm{E}_{\mathrm{K}m}[\mathcal{T}(T)]$ and $\mathrm{E}_{\mathrm{M}}[\mathcal{T}(T)]$ differ greatly.

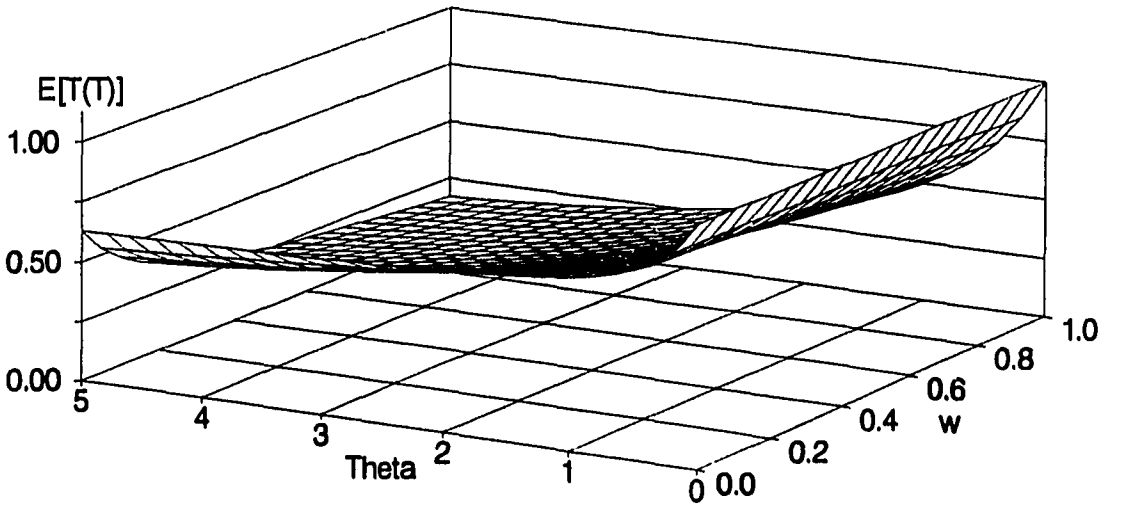


Figure 3.12: $\mathrm{E}_{\mathrm{B}_2}\left[\mathcal{T}_2(T;1,0)|\,\theta,2,w\right]$

Mean timeliness when variance is high

The situation differs for $\varphi > 1$. For one, the Chebyshev upper bound is identically one. If $\varphi >> 1$, only the Markov bound is significant. Note also that

$$\mathrm{E}_{\mathrm{B}r}\left[\mathcal{T}(T)\,|\Phi\right] \;=\; \sum_{j=1}^{r} p_j \mathrm{E}_{\mathrm{M}}\left[\mathcal{T}(T)\,\middle|\theta_j\right] \tag{3.48}$$

is the weighted average of the mean timeliness in the r branches. Because the loss of value is non-linear, $\mathrm{E}_{\mathrm{B}r}[\mathcal{T}(T)]$ can vary greatly for any given θ and $\varphi > 1$. Figure 3.12 illustrates this point by showing $\mathrm{E}_{\mathrm{B}_2}\left[\mathcal{T}_2(T;1,t_0)|\,\theta,2,w\right]$ over the full range of w.

Summary

Mean timeliness depends upon the timeliness model and the distribution of T. If $\mathrm{F}(t)$ is expressed in normalized form, one can show dominance of timeliness models and distributions. In particular, $\mathrm{E}[\mathcal{T}(T)]$ decreases with increasing θ. Although the

Markov lower bound is weak, it does not depend upon φ. The Chebyshev bounds are tighter when φ is small, but become loose with larger φ.

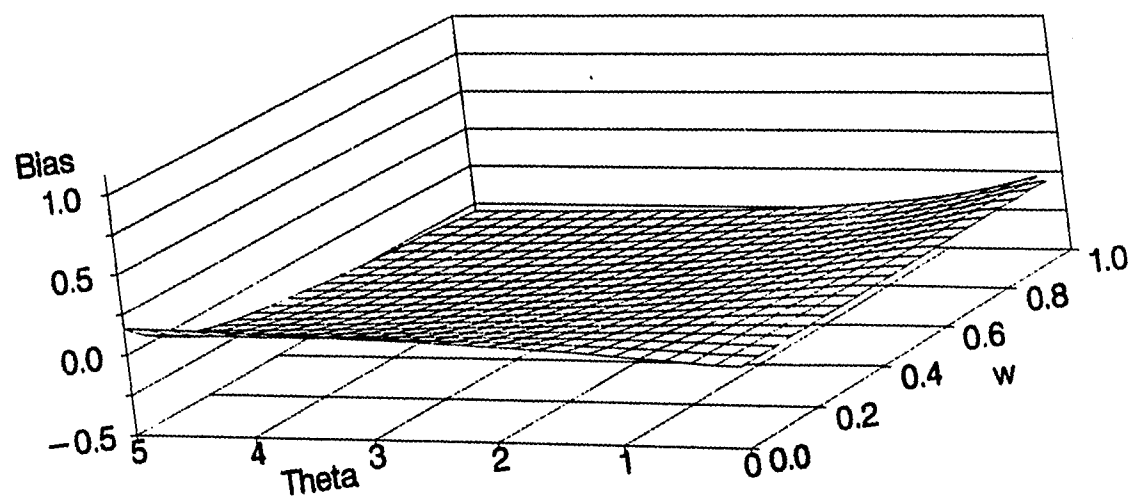
Given θ and $\varphi << 1$, one need not know the exact distribution of T to estimate $E[\mathcal{T}(T)]$. However, when $\varphi > 1$, $E[\mathcal{T}(T)]$ depends greatly on higher moments of T. This implies that no single two-parameter model of $F(t)$ will suffice if $\varphi > 1$.

Robustness of the Hybrid Estimator

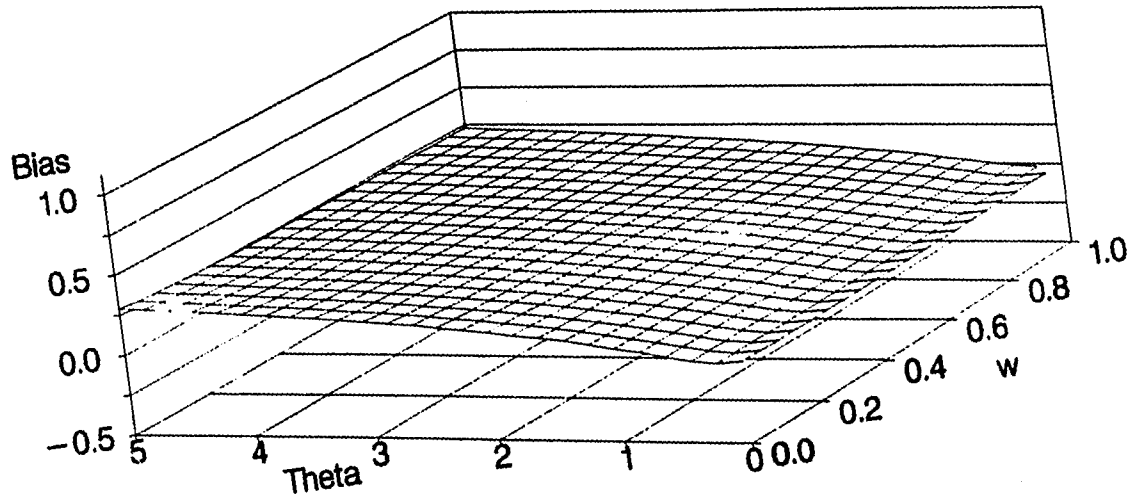
Introduction

This section explores the robustness of the hybrid estimator, $F_H(t)$, as it is used in the proposed method. Robust estimators are estimators for which the range of the bias in an expected range of circumstances is relatively small. Since analysts may not be able to control all circumstances, and often have to model situations in which circumstances vary widely, they would be interested in robust estimators. As mentioned earlier, the common practice of comparing systems on the basis of mean time is not robust. Two-parameter models of $F_T(t)$ perform well when the coefficient of variation is small. However, they may yield misleading results when φ is large. Unless analysts can rule out $\varphi >> 1$, two-parameter estimators are not robust either.

This section first explores the robustness of the hybrid estimator. Then it examines the test for low variance and the likelihood of $\varphi << 1$. Finally, it examines the bias of the proposed method. Briefly, this discussion shows that the proposed method avoids several serious shortcomings of the conventional methods without introducing serious problems of its own.

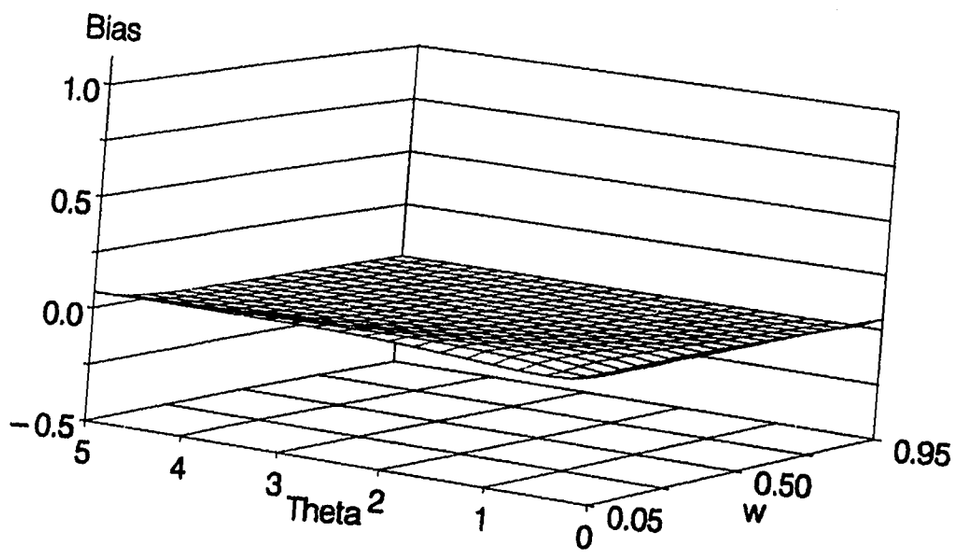


a. $t_0 = 0$

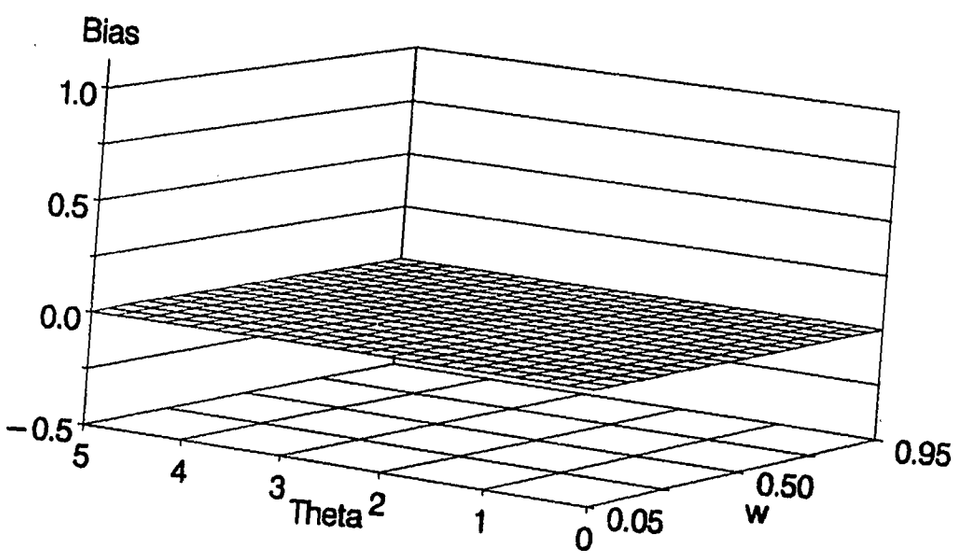


b. $t_0 = 2\tilde{t}$

Figure 3.13: $\mathcal{B}_{\Gamma|B_2}[\mathcal{T}_2(T;1,t_0)|\theta,\varphi]$, the Bias of the Gamma estimate to $E_{B_2}[\mathcal{T}_2(T)]$



a. $t_0 = 0$



b. $t_0 = 2\tilde{t}$

Figure 3.14: The Hybrid Bias $\mathcal{B}_{\text{H}|\mathcal{B}_2}[\mathcal{T}_2(T;1,t_0)|\theta,\varphi]$

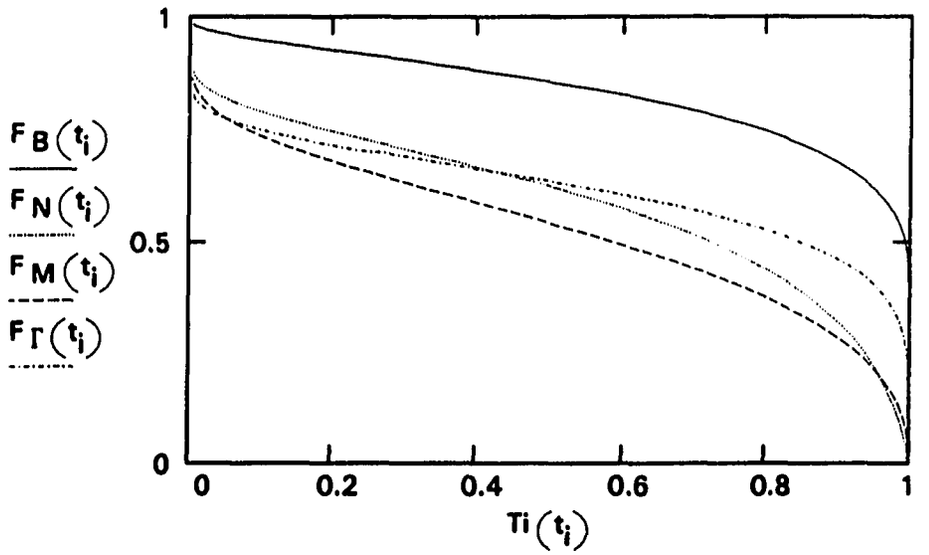


Figure 3.15: Several Approximations to $F_T(\tau)$

Hybrid estimator bias

Earlier comparison of Figures 2.7 and 1.3 implied the hybrid estimator is more robust than other distributions. Figures 3.13 and 3.14 imply the hybrid estimate is more precise than the gamma when $\varphi \geq 1$, but not when φ is very small. Figure 3.15 shows several $F_\Gamma(\tau)$ distributions with equal $\mathcal{T}(t)$, θ, and φ. Figure 3.16 shows two of these $F_\Gamma(\tau)$s together with their hybrid approximations. It implies the hybrid estimator succeeds because it characterizes that part of $F_T(t)$ that has the greatest impact on $E[\mathcal{T}(T)]$. The discussion below is in two parts. The first discusses the bias on $[0, t_\ell)$ and the second the bias on $[t_\ell, \infty)$.

Bias on $[0, t_\ell)$ Consider first bias on the interval $[0, t_\ell)$. Because p_ℓ and $1 - \exp\left(-t_\ell/\theta\right)$ can differ greatly, the Markov bias can be very large. The same

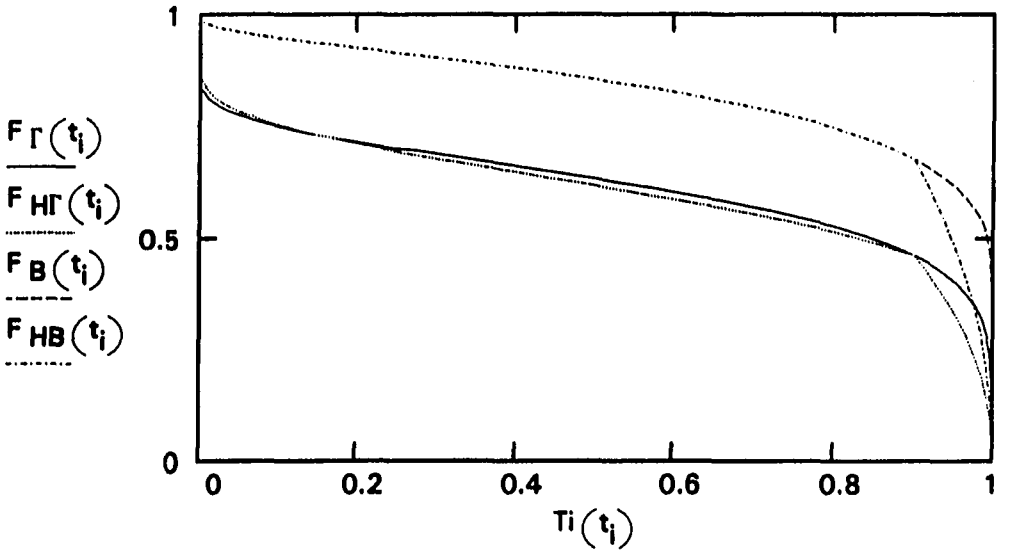


Figure 3.16: The Adaptable Hybrid Approximation of $F_T(\tau)$

condition applies to the gamma and other models based on the first two moments of T. When $\varphi < 1$, the difference between the true and model $F(t_\ell)$ may be small, but when $\varphi > 1$, once again, the difference can be large.

On the other hand, $F_H(t)$ is the straight line connecting $(0,0)$ to (p_ℓ, t_ℓ). The greatest possible value of $E[\mathcal{T}(T)|T < t_\ell]$ is 1.0 and the least is $\mathcal{T}_\ell$. So,

$$\begin{aligned} p_\ell \left\{ \mathcal{T}_\ell - m\left[\mathcal{T}(T)\middle|t_\ell\right] \right\} &\leq \mathcal{B}_{H|T}\left[\mathcal{T}(T)\middle|t_\ell, p_\ell, \Phi_T, T < t_\ell\right] \\ &\leq p_\ell \left\{ 1 - m\left[\mathcal{T}(T)\middle|t_\ell\right] \right\} \end{aligned} \tag{3.49}$$

where $m\left[\mathcal{T}(T)\middle|t_\ell\right]$ is the mean value of $\mathcal{T}(t)$ on $[0, t_\ell)$. The difference between the upper and lower limits is $p_\ell\left(1 - \mathcal{T}_\ell\right)$. So, if $\mathcal{T}_\ell = 0.9$, the maximum uncertainty is about ± 0.05.

Bias on $[t_\ell, \infty)$ Consider next the bias on the upper interval $[t_\ell, \infty)$. At the low end, $\mathrm{F_H}(t_\ell) = p_\ell$, but the difference between $\mathrm{F_T}(t_\ell)$ and either $\mathrm{F_M}(t_\ell)$ or $\mathrm{F_\Gamma}(t_\ell)$ can be very large. If $d\mathcal{T}(t)/dt$ is a decreasing function of T on $[t_\ell, \infty)$, the large difference at t_ℓ can also cause a large bias for $T > t_\ell$. Let $\dot{T} = T - t_\ell$ represent times longer than t_ℓ. The hybrid estimator assumes $\dot{T}$ is an exponential random variable with mean ζ. In addition, $\hat{\zeta}$ ignores the shape of the true distribution beyond t_h. Thus, it ignores the great heaviness in the tails of high-variance distributions, responding instead to the moderate values that affect $\mathrm{E}[\mathcal{T}(T)]$ the most.

The good performance of the hybrid estimator is due partially to its lack of bias at t_ℓ and partially to its emphasis on the critical region $[t_\ell, t_h)$. The strategy works well, as long as the exponential model is a good fit on (t_ℓ, t_h). This is the case when φ is large. For example, suppose $\dot{T}$ follows a B_r distribution.

$$\begin{aligned}
&\mathcal{B}_{\mathrm{H}|\mathrm{B}r}\left[\mathcal{T}(\dot{T} - t_\ell)\,\middle|\,\Phi_{\mathrm{H}} \cup \Phi_{\mathrm{B}r}\right] \\
&\qquad = \sum_{j=1}^{r} p_j\, \mathrm{E_M}\left[\mathcal{T}\left(\dot{T} + t_\ell\right)\middle|\zeta_j\right] \;-\; \mathrm{E_M}\left[\mathcal{T}\left(\dot{T} + t_\ell\right)|\zeta\right] \\
&\qquad = \sum_{j=1}^{r} p_j \left\{\mathrm{E_M}\left[\mathcal{T}\left(\dot{T} + t_\ell\right)\middle|\zeta_j\right] \;-\; \mathrm{E_M}\left[\mathcal{T}\left(\dot{T} + t_\ell\right)|\zeta\right]\right\} \qquad (3.50)
\end{aligned}$$

is the weighted average of exponential expectations. Those branches with $0 < \zeta_j < t_h - t_\ell$ will have the greatest influence on both ζ and $\mathrm{E}[\mathcal{T}(T)]$. This also means the significant ζ_j will be close to ζ in value, and regardless of the size of p_j, will contribute little to bias. The term error will be large when $\zeta_j >> \zeta$, but in these cases, p_j will be small, so again there is little contribution to error.

However, when φ is small, the situation changes. For example, assume $\dot{T}$ follows

a K_m distribution. As above,

$$\mathcal{B}_{\mathrm{H}|\mathrm{K}m}\left[\mathcal{T}(\dot{T}+t_\ell)\,\middle|\,\Phi_{\mathrm{H}}\cup\Phi_{\mathrm{K}m}\right] =$$

$$\sum_{i=1}^{m} A_i\left(\Phi_{\mathrm{K}m}\right)\left\{\mathrm{E}_{\mathrm{M}}\left[\mathcal{T}\left(\dot{T}+t_\ell\right)|\zeta_j\right] - \mathrm{E}_{\mathrm{M}}\left[\mathcal{T}\left(\dot{T}+t_\ell\right)|\zeta\right]\right\}. \quad (3.51)$$

Once again, the significant ζ_j will be close to each other in value, but the A_i alternate in sign and can be very large. This can lead to a large estimation error.

The worst case occurs when $\dot{T}$ has very low variance. In this case, $\mathrm{F_T}(\dot{t})$ approaches zero for $\dot{t} < \zeta$ and one for $\dot{t} > \zeta$. Thus, the integral on $[t_\ell, \zeta + t_\ell)$ has its smallest value and that on $[\zeta + t_\ell, \infty)$ its highest. If the majority of the value is in the smaller interval, the bias will become very negative. If the majority of value is in the larger interval the bias will become very positive. If $\dot{T} \equiv \zeta$, then $\mathrm{E}\left[\mathcal{T}(\dot{T}+t_\ell)\right] = \mathcal{T}(\zeta + t_\ell)$. The hybrid estimate would be $\mathrm{E_M}\left[\mathcal{T}(\dot{T}+t_\ell)\,|\zeta\right]$. So, as shown in Figure 3.14, when $\varphi \approx 0$, the hybrid bias is essentially that of the Markovian estimate.

The total bias in the extreme case $T \equiv \theta$ is analyzed below. First, $\mathrm{E}[\mathcal{T}(T)|\theta] = \mathcal{T}(\theta)$. The hybrid parameters are:

$$p_l = \begin{cases} 1 & , \theta \le t_\ell \\ 0 & , \theta > t_\ell \end{cases} \quad (3.52)$$

and

$$\zeta \quad \begin{cases} = \quad 0 & , \theta \le t_\ell \\ = \quad \theta - t_\ell & , t_\ell < \theta \le t_h \\ > \quad t_h - t_\ell & , \theta > t_h. \end{cases} \quad (3.53)$$

The implied extreme limits for the $\mathcal{T}_2(t)$ model are summarized in Table 3.5.

Table 3.5: Asymptotic Limits on $\mathcal{B}_{\mathrm{H}|\mathrm{K}m}[\cdot]$ with $\mathcal{T}_\ell = 0.9$ and $\mathcal{T}_h = 0.1$

Value of θ	Limits on $\mathcal{B}_{\mathrm{H}\|\mathrm{K}m}[\mathcal{T}_2(T)]$		
	Lower	Upper	Range
$0 \le \theta \le t_0$	$1.0 - C_1$		0
$t_0 < \theta \le t_\ell$	$0.9 - C_1$	$1.0 - C_1$	0.1
$t_\ell < \theta \le t_h$	−0.1833	0.1	0.283
$t_h < \theta$	0	0.1	0.1
$t_\ell < \theta < \theta_{0.5}$	−0.097	0.1	0.192

A decision rule The hybrid estimate is not robust when $t_\ell < \theta < t_h$ and φ is small. If the output is uncorrelated, one can simply estimate φ. However, under the best conditions, $\hat{\varphi}$ would be noisy, and the output is almost always autocorrelated.

One indication of trouble with the hybrid estimate is that $\mathrm{F}_\mathrm{T}(T)$ rises quickly. In the extreme deterministic case, $\theta \in \Omega$ implies $p_h - p_\ell = 1$. Figure 3.17 shows $p_h - p_\ell$ as a function of θ and φ, assuming a gamma distribution. Here $t_\ell/\theta = 1$ and $t_h/\theta = 2$. The pattern is similar for other distributions and other choices of t_ℓ and t_h. One can think of other situations as independent re-scaling within each of $[0, t_\ell)$, $[t_\ell, t_h)$, and $[t_h, \infty)$. There is typically a sharp rise in value for points in the triangular region approximately bounded by $(\theta, \varphi) = (t_\ell, 0)$, $(t_h, 0)$, and $\left(\frac{t_\ell + t_h}{2}, 0.5\right)$.

In practice, one usually need not test for this situation. First of all, $p_\ell > 0.5$ is evidence that $p_h - p_\ell < 0.5$. Secondly, Figure 3.18 shows φ as a function of ρ in the (M/D/1) model. The variance of the service distribution is zero and it is unlikely the arrival distribution in $\mathrm{C}^3\mathrm{I}$ systems will be much less variable than the exponential. Note that unless $\rho < 0.5$, $\varphi > 0.5$, implying that low-variance situations are rare in

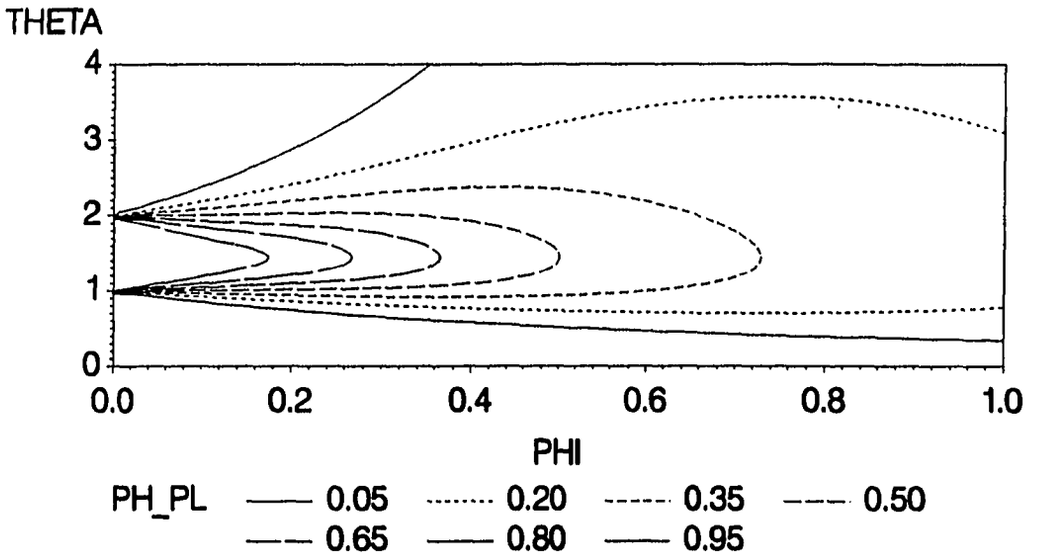


Figure 3.17: The Difference $p_h - p_\ell$ when T is Distributed as a Gamma Random Variable

heavily-loaded systems.

The bias of the proposed method

The method depends upon four key points. First, preliminary screening eliminates regions in which $\mathrm{E}[\mathcal{T}(T)]$ will be too small. In these regions, even if the error is large, it will be unimportant. Second, focused means and probability estimates are usable over a very wide set of circumstances. Third, the decision rule determines when to use the hybrid model, and fourth, situations in which $\mathrm{F_H}$ are inappropriate are rare.

Figure 3.19 illustrates typical regions in the low-variance segment of the $\theta \times \varphi$ plane. The pattern extends smoothly into the high-variance region, with the pre-screening upper limit becoming larger with φ. Figure 3.20 illustrates the net effect

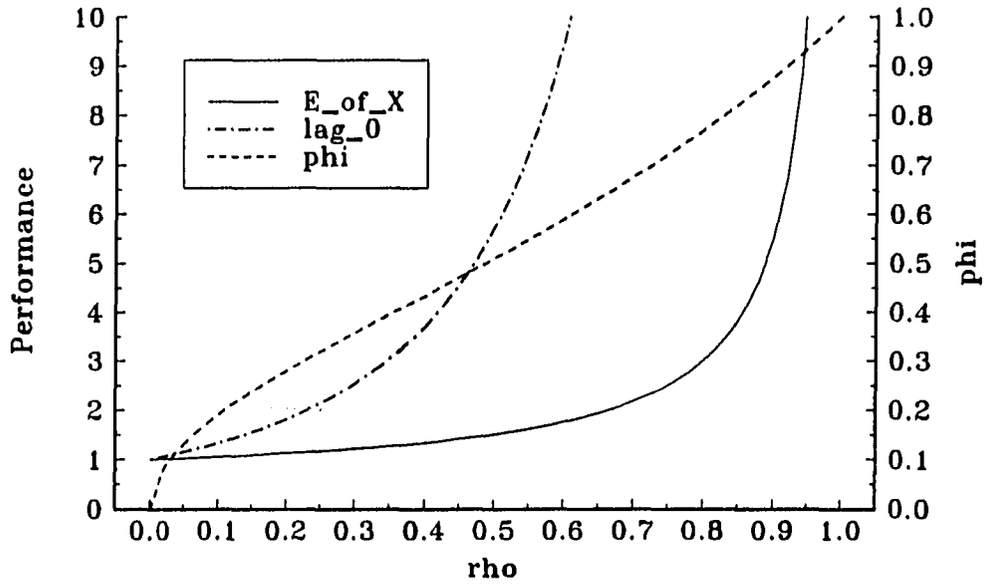


Figure 3.18: Characteristics of the (M/D/1) Queue as a Function of ρ for $\theta = 1$

of the entire method on bias in a number of situations. All plots are of $\mathrm{E_H}\left[\mathcal{T}_2(T)\right]$ with $\mathcal{T}_\ell = 0.9$ and $\mathcal{T}_h = 0.1$. They differ in the underlying distribution. To provide visual continuity, $\mathcal{B}_{\mathrm{H}}[\cdot]$ was set to zero in those instances in which $\mathrm{E_H}[\mathcal{T}(T)]$ would not be computed.

Conclusion

This exploration brings out three important points about situations in which timeliness is a non-linear function of time. First of all, one needs to consider the shape of the distribution of T more carefully than is common practice. Secondly, because of the much greater difficulty of estimating shape elements of a distribution, one should focus on the regions of greatest impact on the effectiveness measure.

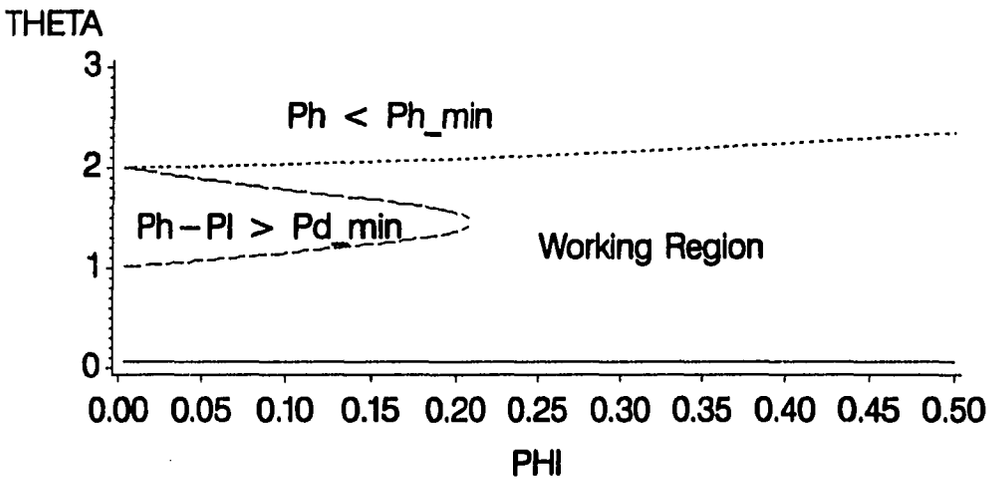
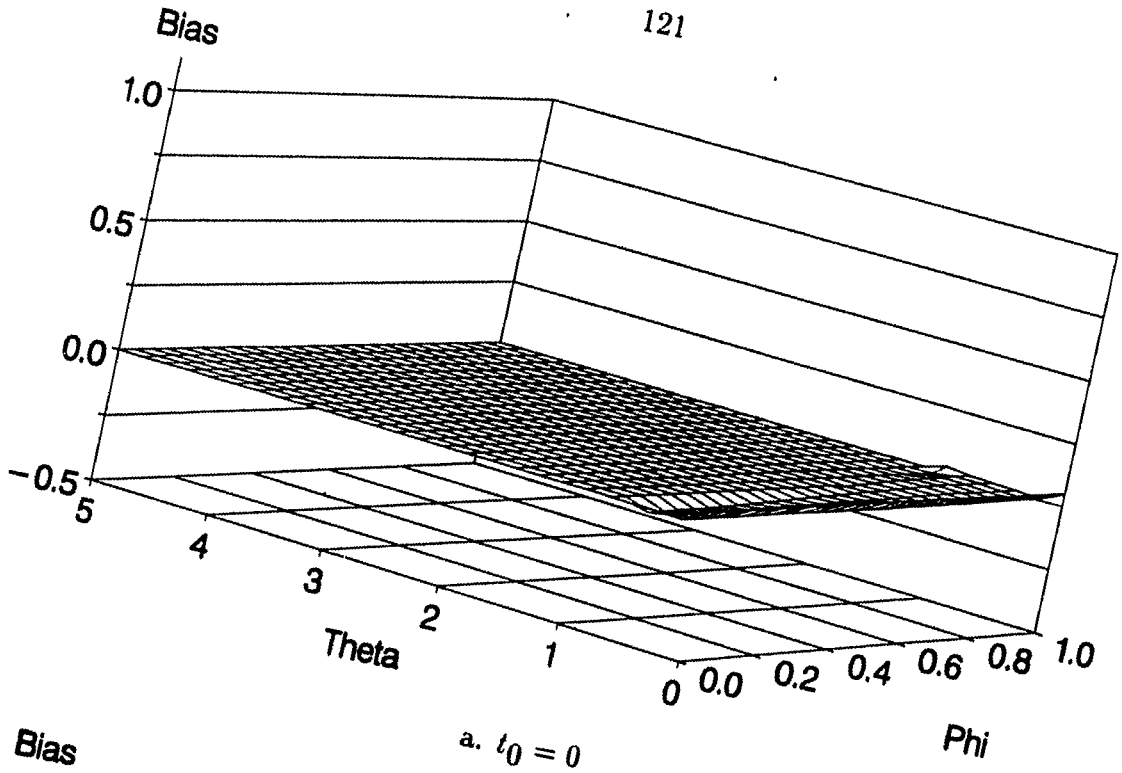


Figure 3.19: Significant Regions for the Proposed Method

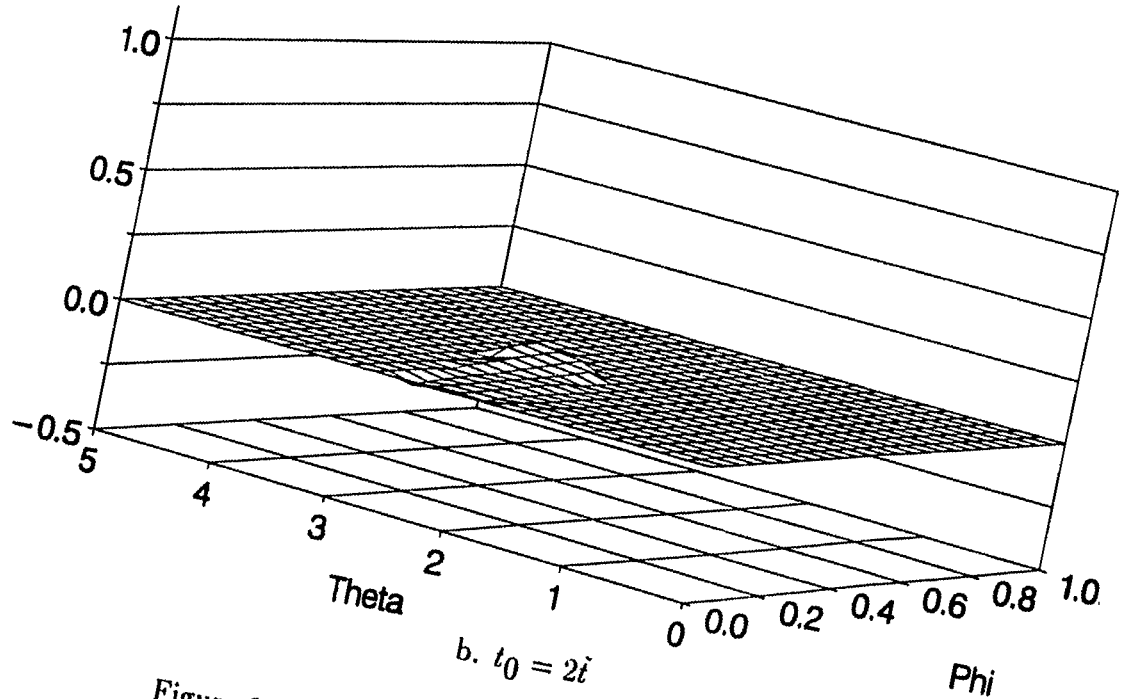
Finally, since the relation of effectiveness to time may not be exact, one needs an adaptive method that allows sensitivity analysis.

Although using unfocused means is a low-cost simple method, is not good practice in these cases. First, the analyst is likely to miss-classify systems, unless the distribution of T does not differ among candidate models. Secondly, if the distribution is not approximately exponential, $\hat{\mathrm{E}}_{\mathrm{M}}[\mathcal{T}(T)]$, or any other estimate based on $\mathrm{E}[T]$, will not be a good predictor of system performance. Finally, the use of such models creates a body of performance measures that are not comparable across different studies.

The use of a conventional distribution, such as the gamma, would be an improvement, if one could estimate φ. However, simulation output is often autocorrelated. Additionally, when $\varphi > 1$, higher moments of $\mathrm{F}(t)$ can greatly affect $\mathrm{E}[\mathcal{T}(T)]$. These



a. $t_0 = 0$



b. $t_0 = 2\tilde{t}$

Figure 3.20: Effective Bias of the Proposed Method

weaknesses are shared by all models whose parameters are based on the first two moments of T.

Because it focuses on $\Omega = (t_\ell, t_h)$, the interval of interest, the hybrid estimate performs much better than either of the two more common approaches. Its extreme limits on bias are reasonably small. However, its bias in practice is much smaller, when used in the context of the proposed method to analyze moderately- to heavily-loaded C^3I systems.

CHAPTER 4. CONCLUSION

Summary

This paper stresses four major themes. First, the effect of delay upon value is a non-increasing non-linear function of time. This implies that mean time is not a good predictor of time effectiveness. The expectation of timeliness is presented as a superior measure that is simply related to time effectiveness. Second, the development of C^3I systems requires the concerted effort of specialists from numerous fields. This implies a need for inter-disciplinary communication. The Focused Measure of Performance is presented to improve such communication. Third, analysts need to consider the effects of uncertainty in the description of offered loads, environments, and effectiveness models in judging the quality of their answers. The common practice of specifying C^2 values as exact points fools us into thinking we know more about the system than we do. The pseudo-confidence interval, together with an extension of Taguchi's approach that allows for uncertainty in the definition of quality is presented as a quantitative method to gauge the precision of estimates. Forth, one may use the focused measure of performance and pseudo-confidence intervals to find productive areas to work in across the C^2C^3I barrier. Finally, C^3I systems are queueing systems by nature, and as such have moderate to high autocorrelation. This autocorrelation complicates the problem of determining the distribution of T. Several bounds and

estimators of mean timeliness are presented that are both practical in the presence of autocorrelation and a vehicle to good estimates of mean timeliness.

This report proposes a philosophy of analysis, a measure of performance, a measure of effectiveness, and a method of analysis that simultaneously addresses these themes. Chapter 1 points out the fundamental problems estimating C^3I using conventional measures of performance, such as the mean and standard deviation. It points out the complications due to autocorrelation and unknown certainty of modeling data. The proposed method, presented in Chapter 2, attempts to alleviate these difficulties by using a context-sensitive performance measure, rather than an unconditional one, in hopes of obtaining better approximations of effectiveness. The method also advocates stating the uncertainty in modeling and evaluation data and carrying that uncertainty forward through a modification of Taguchi's experiment designs and pseudo-confidence intervals. Finally, Chapter 3 discusses the ability of the proposed method to work under a variety of circumstances. These chapters show the need for and a means of improving communications across the C^2–C^3I interface. They also show the proposed method is 1) practical, 2) robust, 3) efficient, and 4) an improvement over many conventional methods.

Conclusions

Analysts need to consider four important facts when evaluating time performance. First, because timeliness is non-linear, they cannot base timeliness measures upon $E[T]$. Second, except in very simple cases, they cannot use existing analytical network models to determine $F(t)$. Third, because of autocorrelation, they cannot estimate $F(t)$, $f(t)$, or $h(t)$ directly from simulation output. Finally, because ana-

lysts do not know exactly the input, working conditions, and MOP/MOE conversion functions, they must not only use methods that are robust over a wide range of conditions, but also consider the uncertainty in this data when evaluating results. The proposed method is one way to addresses these concerns.

Recommendations for Further Study

The approach used in this paper is simple, but effective. Although this demonstration shows that the concept is workable, better methods could be developed. Now that there is a way, the analytical community can seek a better one. Further work, however, will partially depend upon increased C^2 knowledge.

For this work to be of lasting value, the C^2 community needs to address two problems. First, although the CDB represents a great deal of effort and is very valuable, it could be improved. The uncertainty in COF codes and some minor issues about perishability codes need to be resolved. In addition, the C^2 community should agree on some minimal characterization of timeliness. This can be as simple as specifying two value points, e.g. $\left(t_\ell, p_\ell\right)$ and $\left(t_h, p_h\right)$, or as ambitious as a generic $\mathcal{T}(t)$ model.

Most of C^3I analysis deals only with point estimates, or at most, intervals based on the variation within simulation runs. Because these have an unknown degree of precision, the analyst cannot know the quality of such estimates. Unless the analyst takes into account external uncertainties in problem formulation and performance evaluation, such estimates only reflect their ability to predict the performance of the C^3I model, rather than the C^3I system it represents. Therefore, the analyst must also conduct a sensitivity analysis of the results to these uncertainties. This in

turn implies that the C^2 community needs to more accurately describe the existing variation in its estimates of offered load, environmental conditions, and effectiveness models.

Finally, although the focus of this paper is the effect of delay on telecommunications in C^3I systems, most of the discussion is applicable to the effects of delay or storage in any time-critical situations. Thus, another area of research is to determine the similarities and differences in similar situations.

BIBLIOGRAPHY

[1] S. J. Andriole and S. M. Haplin, "Information technology for command and control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-16, pp. 762–765, November/December 1986.

[2] M. Athans, "Command and control (C2) theory: a challenge to control science," *IEEE Transactions on Automatic Control*, vol. AC-32, pp. 286–293, April 1987.

[3] Naval Ocean Systems Center, "Handbook of C^3 theory – appendix to HFOSL technical note 41-87-01," Tech. Rep., Naval Ocean Systems Center, San Diego, CA, 1986.

[4] P. A. Kennedy, "Command and control in the international arena," *International Conference on Advances in Command, Control and Communications Systems*, pp. 164–168, April 1985.

[5] T. King and A. F. Martin, "The modeling of communications networks," *International Conference on Advances in Command, Control and Communications Systems*, pp. 90–98, April 1985.

[6] D. N. Kyle, "A methodology for air command and control system (ACCS) design," *International Conference on Advances in Command, Control and Communications Systems*, pp. 86–89, April 1985.

[7] D. J. Morris, *Introduction to Communication Command and Control Systems.* New York: Pergamon Press, second ed., 1983.

[8] I. White, "The future of command systems," in *Advances in Command, Control & Communication Systems*, (C. J. Harris and I. White, eds.), ch. 1, pp. 1–15, London, UK: Peter Peregrinus Ltd., 1985.

[9] A. H. Levis and M. Athans, "The quest for a C^3 theory: dreams and realities," in *Science of Command and Control: Coping with Uncertainty*, (S. E. Johnson and A. H. Levis, eds.), ch. 2, pp. 4–9, Washington, DC: AFCEA International Press, 1988.

[10] I. White, "Raising standards in command, control and communications," *Proceedings of the Third International Conference on Command, Control, Communications and Management Information Systems*, pp. 184–191, May 1989.

[11] R. Sweet, M. Metersky, and M. Sovereign, "Command and control evaluation workshop," Tech. Rep., MORS C^2 MOE Workshop, Naval Postgraduate School, January 1985.

[12] R. Sweet, D. Mensh, P. Gandee, I. Stone, and K. Briggs, "The modular command and control evaluation structure (MCES): applications of and expansion to c3 architectural evaluation," Tech. Rep., Naval Postgraduate School, September 1986.

[13] W. Kuo, J. Mullen, J. Rupe, and S. Gavirneni, "Command control and communication information network analysis tool (C3INAT): final report," Tech. Rep., Prepared for the U. S. Army Communications-Electronics Command and Fort Monmouth, NJ., November 1990.

[14] Teledyne-Brown Engineering, "USASC CDSF communications data base (CDB) phase III (with software documentation)," Tech. Rep., Teledyne-Brown Engineering, Augusta, GA, March 1989.

[15] C. G. Q. Coe and D. J. T. Dockery, "OJCS initiatives in C^2 analysis and simulation," in *Science of Command and Control: Coping with Uncertainty*, (S. E. Johnson and A. H. Levis, eds.), ch. 4, pp. 19–31, Washington, DC: AFCEA International Press, 1988.

[16] D. K. Hitchens, "A general theory of command and control," *Proceedings of the Third International Conference on Command, Control, Communications and Management Information Systems*, pp. 111–126, May 1989.

[17] A. L. Lopez, "A study in the diversity in methodologies for the derivation of measures of effectiveness and performance," in *Science of Command and Control: Coping with Uncertainty*, (S. E. Johnson and A. H. Levis, eds.), ch. 12, pp. 100–105, Washington, DC: AFCEA International Press, 1988.

[18] I. Mayak and I. Rubin, "Paradigms for understanding C^3, anyone?," in *Science of Command and Control: Coping with Uncertainty*, (S. E. Johnson and A. H. Levis, eds.), ch. 6, pp. 48–61, Washington, DC: AFCEA International Press, 1988.

[19] D. R. Sweet, "The MCES and the search for generic measures," in *Science of Command and Control: Coping with Uncertainty*, (S. E. Johnson and A. H. Levis, eds.), ch. 13, pp. 106–117, Washington, DC: AFCEA International Press, 1988.

[20] A. E. R. Woodcock, "Toward a theory of combat with embedded command and control," *Proceedings of the Third International Conference on Command, Control, Communications and Management Information Systems*, pp. 127–138, May 1989.

[21] W. R. Beam, *Command, Control, and Communications Systems Engineering.* New York: McGraw-Hill, 1989.

[22] K. E. E. Raatikainen, "Approximating response time distributions," *Performance Evaluation Review*, vol. 17, pp. 190–199, May 1989.

[23] T. M. Apostol, *Mathematical Analysis.* Reading, MA: Addison-Wesley, 2nd ed., 1974.

[24] P. H. Cothier and A. Levis, "Timeliness and measures of effectiveness in command and control," *IEEE Transactions on Systems*, vol. SMC-16, pp. 844–853, November/December 1986.

[25] P. J. Burke, "The output of a queueing system," *Operations Research*, vol. 4, pp. 699–704, 1956.

[26] J. A. White, J. W. Schmidt, and G. K. Bennett, *Analysis of Queueing Systems.* New York, NY: Academic Press, 1975.

[27] J. Kleijnen, *Statistical Tools for Simulation Practitioners.* New York: Marcel Dekker, Inc, 1987.

[28] J. F. Lawless, *Statistical Models and Methods for Lifetime Data.* New York: John Wiley & Sons, 1982.

[29] J. E. Freund and R. E. Walpole, *Mathematical Statistics.* Englewood Cliffs, NJ: Prentice-Hall, 3rd ed., 1980.

[30] D. G. Luhenberger, *Introduction to Linear and Nonlinear Programming.* Reading, MA: Addison-Wesley, 1965.

[31] G. Taguchi, E. A. Elsayed, and T. Hsiang, *Quality Engineering in Production Systems.* New York, NY: McGraw-Hill Book Company, 1989.

[32] G. Taguchi, *Introduction to Quality Engineering.* Available from Unipub Quality Resources, White Plains, NY: Aisian Productivity Organization, 1986.

[33] G. Taguchi and Y. Wu, *Introduction to Off-Line Quality Control.* Meieki Nakamura-Ku Nagaya, Japan: Central Japan Quality Control Association, 1985.

[34] G. Box, S. Bisgaard, and C. Fung, "An explanation and critique of Taguchi's contributions to quality engineering," *Quality and Reliability Engineering International*, vol. 4, pp. 123–131, April-June 1988.

[35] B. Gunter, "A perspective on the Taguchi methods," *Quality Assurance*, vol. 13, pp. 81–87, September 1987.

[36] N. Harris, "Editorial," *Quality and Reliability Engineering International*, vol. 4, pp. 85–86, April–June 1988.

[37] R. N. Kackar, "Off-line quality control, parameter design, and the Taguchi method," *Journal of Quality Technology*, vol. 17, pp. 176–188, October 1985.

[38] C. R. Hicks, *Fundamental Concepts in the Design of Experiments.* New York, NY: Holt, Rinehart and Winston, second ed., 1973.

[39] A. Gafarian, C. Ancker, Jr., and T. Morisaku, "Evaluation of commonly used rules for detecting steady state in computer simulation," *Nav. Res. Log. Q.*, vol. 25, pp. 511–529, 1978.

[40] J. Wilson and A. Pritsker, "A survey of research on the simulation startup problem," *Simulation*, vol. 31, pp. 55–58, August 1978.

[41] J. R. Wilson and A. B. Pritsker, "Evaluation of startup policies in simulation experiments," *Simulation*, vol. 31, pp. 79–89, September 1978.

[42] B. Pagurek and C. M. Woodside, "The sum of serial correlations of waiting and system times in $Gi/G/1$ queues," *Operations Research*, vol. 27, pp. 755–766, July-August 1979.

[43] J. Kleijnen, *Statistical Techniques in Simulation, Part II.* New York: Marcel Dekker, Inc, 1985.

[44] N. L. Johnson and S. Kotz, *Continuous Univariate Distributions - 1.* New York, NY: John Wiley & Sons, 1970.

[45] S. M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists.* New York, NY: John Wiley & Sons, 1987.

[46] D. Gross and C. Harris, *Fundamentals of Queueing Theory.* New York: John Wiley & Sons, 2nd ed., 1985.

[47] L. Kleinrock, *Queueing Systems.* Vol. I: Theory, New York: John Wiley & Sons, 1975.

[48] W. L. Smith, "On the distribution of queueing times," *Proceedings of the Cambridge Philosophical Society,* vol. 49, pp. 449–461, 1953.

[49] H. Scheffé, "A useful convergence theorem for probability distributions," *Annals of Mathematical Statistics,* vol. 18, pp. 434–438, September 1947.

[50] V. I. Pagurova, *Tables of the Exponential Integral* $E_\nu(x) = \int_1^\infty e^{-xu} u^{-\nu} du$. New York: Pergamon Press, 1961.

[51] D. J. Daley, "The serial correlation coefficients of waiting times in a stationary single server queue," *Journal of the Australian Mathematical Society,* vol. 8, pp. 683–699, 1968.

[52] P. M. Morse, "Stochastic properties of waiting lines," *Journal of the Operations Research Society of America,* vol. 3, pp. 255–261, August 1955.

[53] A. G. Pakes, "The serial correlation coefficients of waiting times in the stationary $Gi/M/1$ queue," *The Annals of Mathematical Statistics,* vol. 42, no. 5, pp. 1727–1734, 1971.

[54] C. M. Woodside and B. Pagurek, "An algorithm for computing serial correlations of times in Gi/G/1 queues with rational arrival processes," *Management Science,* vol. 25, pp. 54–63, January 1979.

[55] P. Welch, "The statistical analysis of simulation results," in *The Computer Performance Modeling Handbook,* (S. Lavenberg, ed.), ch. 6, pp. 268–328, Academic Press, 1983.

[56] P. A. Lewis and E. J. Orav, *Simulation Methodology for Statisticians, Operations Analysts, and Engineers.* Pacific Grove, CA: Wadsworth & Brooks/Cole, 1989.

[57] G. S. Fishman, "Grouping observations in digital simulation," *Management Science*, vol. 24, pp. 510–521, January 1978.

[58] L. W. Schruben, "Detecting initialization bias in simulation output," *Operations Research*, vol. 30, no. 3, pp. 569–590, 1982.

[59] R. Roy, *A Primer on the Taguchi Method.* New York, NY: Van Nostrand Reinhold, 1990.

[60] E. S. Pearson and H. O. Hartley, "Charts of the power function for analysis of variance tests, derived from the non-central f-distribution," *Biometrika*, vol. 38, no. 1 & 2, pp. 112–130, 1951.

ACKNOWLEDGEMENTS

I am deeply grateful to Dr. Herbert T. David for contributing his valuable time and insights as well as encouraging me to do my very best in this project. I also thank Dr. Roger W. Berger, who encouraged me to find balance in this endeavor. I thank Dr. Way Kuo, who introduced me to C^2 and C^3I problems. I also appreciate the contributions of Dr. Richard Tondra, Dr. Doug Gemmill, and Dr. Peter Sherman, who served on my committee and offered guidance on selected topics.

I wish to thank my employer, Dr. Satish Kamat, who allowed me the time and resources I needed in this effort. I also thank my family, for their forbearance over the course of this study. I extend my highest admiration and gratitude to my wife, Patricia, who exhibited all of those fine characteristics for which I love her deeply.

Finally, I dedicate this work to the memory of Dr. Vince Sposito, a former member of my committee and a great inspiration to me and many others.

APPENDIX A. DETERMINING CDB TIMELINESS PARAMETERS

The following discusses adjustments made to CDB data and determines parmeter values from the CDB codes. It assumes that $T(t_\ell) \stackrel{\Delta}{=} \mathcal{T}_\ell = 0.9$ and $T(t_h) \stackrel{\Delta}{=} \mathcal{T}_h = 0.1$.

CDB Code Adjustments

The CDB perishability boundaries are not consistently defined. In most cases, t_ℓ for $\mathcal{P}$ range coincides with t_h for $\mathcal{P}+1$, but in some it does not. The analyst will be using simulation to estimate probabilities, so if one time is both a t_ℓ and a t_h, an estimate of p_ℓ for code $\mathcal{P}$ will be the same as that of p_h for code $\mathcal{P}+1$. the entire range of $\mathcal{P}$ codes could be represented with fourteen specific times instead of eighteen. The problem is especially acute for $5 \leq \mathcal{P} \leq B$. To solve this problem, let $t_h(\mathcal{P}) = t_\ell(\mathcal{P}-1)$ for $\mathcal{P} > 0$. Since this rule changes only the larger of the two limits, it will lead to the least change in values.

Since T is a continuous random variable, there is no difference between $\Pr\{T \leq t\}$ and $\Pr\{T < t\}$. However, digital computers deal with a subset of the rational numbers, so one must adapt a convention for coincidence. Let these pairs represent a series of half-open intervals $[t_\ell, t_h)$, so if $T = t_\ell(\mathcal{P})$ it will be counted in the $\mathcal{P}$-th interval, but if $T = t_h(\mathcal{P})$, it will be counted in the $(\mathcal{P}+1)$-th. These modified ranges mean that if the analyst wishes to examine R contiguous ranges, they will only need

to estimate $R + 1$ probabilities.

Parameter Determination

Exponential timeliness model

Let $\mathcal{T}_1(t;\tilde{t}) = \exp(-t/\tilde{t})$. In general, one cannot find a $\tilde{t}$ that will cause $\mathcal{T}_1(t;\tilde{t})$ to pass through both constraining points. However, the analyst can use a regression model, so

$$\tilde{t} = \frac{t_\ell + t_h}{-\left(\ln \mathcal{T}_\ell + \ln \mathcal{T}_h\right)}. \tag{A.1}$$

One can convert most CDB perishability codes in this manner. However, when $\mathcal{P} = 0$, $t_h = +\infty$, so one must either rely on $(t_\ell, \mathcal{T}_\ell)$ alone or specify a finite t_h. The largest ratio of t_h to t_ℓ among the other codes in the hour range is 2.0. Let $t_h = 24$ hours when $\mathcal{P} = 0$.

Two-parameter exponential model

Of course, $\mathcal{T}_\ell \leq I$, implies $t_0 \leq t_\ell$. Substituting both constraining points into the lower definition of (1.7), taking logarithms and solving the two simultaneous equations yields:

$$\tilde{t} = -\frac{t_h - t_\ell}{\left(\ln \mathcal{T}_h - \ln \mathcal{T}_\ell\right)} \tag{A.2}$$

and

$$t_0 = \frac{t_\ell \ln \mathcal{T}_h - t_h \ln \mathcal{T}_\ell}{\ln \mathcal{T}_h - \ln \mathcal{T}_\ell}. \tag{A.3}$$

This method allows one to convert most of the CDB perishability codes, handling $\mathcal{P} = 0$ as above. However, when $\mathcal{P} = B$, $t_\ell = 0$, which would imply a negative

t_0. Instead of this, let $t_{0_B} = 0$. This leads to $\tilde{t}_B = -\frac{t_{h_B}}{\ln \mathcal{T}_h - \ln \mathcal{T}_\ell}$. Then, since $\mathcal{T}(t_\ell) = \mathcal{T}_\ell$, $t_{\ell_B} = -\tilde{t}_B \ln \mathcal{T}_\ell$. This preserves the meanings of t_ℓ and t_h as $\mathcal{T}^{-1}(\mathcal{T}_\ell)$ and $\mathcal{T}^{-1}(\mathcal{T}_h)$, respectively, and avoids the difficulties that arise if $t_0 < 0$.

Weibull model

The Weibull survivor function is commonly used to represent residual value. Let

$$\mathcal{T}_{\mathrm{W}}(t; \alpha, \beta) \triangleq \exp\left[-(t/\beta)^\alpha\right]. \tag{A.4}$$

Substituting the constraining points into (A.4), taking logarithms twice and solving the simultaneous equations leads to:

$$\alpha = \frac{\ln\left(-\ln \mathcal{T}_h\right) - \ln\left(-\ln \mathcal{T}_\ell\right)}{\ln t_h - \ln t_\ell} \tag{A.5}$$

and

$$\beta = \exp\left[\frac{\ln t_\ell \ln\left(-\ln \mathcal{T}_h\right) - \ln t_h \ln\left(-\ln \mathcal{T}_\ell\right)}{\ln\left(-\ln \mathcal{T}_h\right) - \ln\left(-\ln \mathcal{T}_\ell\right)}\right] \tag{A.6}$$

One can convert the CDB perishability codes, handling $\mathcal{P} = 0$ and $\mathcal{P} = B$ as above.

A Comparison of the Models

Given enough (t_i, v_i) pairs, one can check the fit of each model to the data and select the most representative. However, because the CDB contains only two points, one must decide among candidate models on some other basis.

Figure A.1 illustrates plots of the three functions for $\mathcal{P} = 5$. The one-parameter exponential shows a very rapid initial decline in value and has a value of only about 0.7, rather than 0.9, at t_ℓ. The two-parameter exponential model passes through

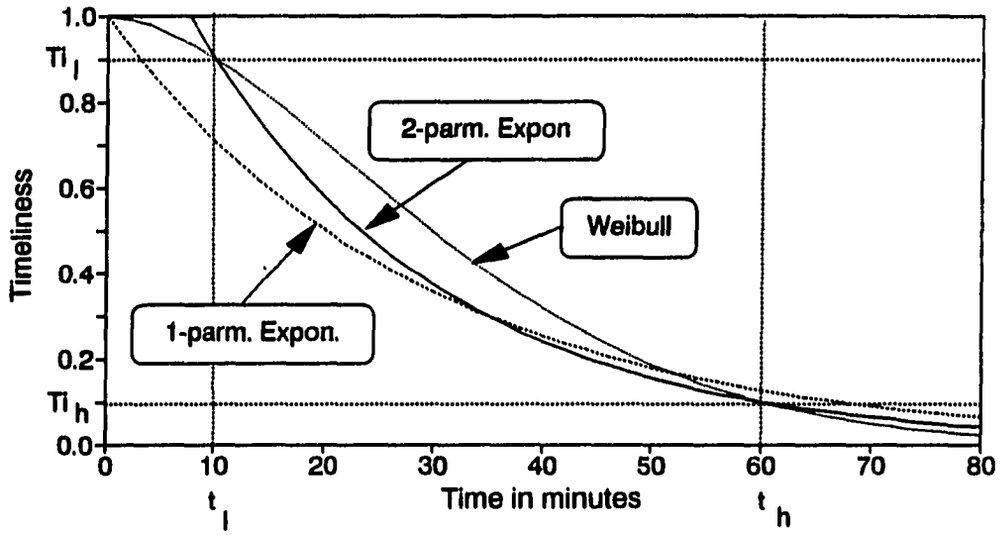


Figure A.1: A Comparison of Three Timeliness Models for $\mathcal{P} = 5$

$(t_\ell, \mathcal{T}_\ell)$ and $(t_h, \mathcal{T}_h)$ and predicts a rapid loss in value near t_ℓ. The Weibull model also passes through both constraining points, but predicts a more gradual loss in value over (t_ℓ, t_h).

The Weibull shape parameter (α), which ranges from about one to eleven over the $\mathcal{P}$ codes, allows the analyst to choose the region of greatest change in $\mathcal{T}(t)$. However,

$$\frac{d\mathcal{T}_{\mathrm{W}}(t;\alpha,\beta)}{dt} = -\alpha\left(\frac{t}{\beta}\right)^{\alpha-1}\exp\left[-\left(\frac{t}{\beta}\right)^{\alpha}\right]. \qquad \text{(A.7)}$$

Since all CDB vlaues of α exceed 1.0, the value will drop slowly for small values of t and the Weibull model will predict a more gradual loss in value near t_ℓ than the two-parameter exponential. Thus, the two-parameter exponential model seems to model probable value loss better than the Wiebull.

Table A.1: Reference Times for $\mathcal{T}_2(t;\tilde{t},t_0)$

$\mathcal{P}$	Units	t_ℓ ($t_{0.9}$)	t_h ($t_{0.1}$)	t_0	$\tilde{t}$	t_F	$t_{0.99}$	$t_{0.01}$
0	Hrs.	8	24	7.233	7.282	16.	7.306	40.767
1	Hrs.	4	8	3.808	1.820	4.	3.826	12.192
2	Hrs.	3	4	2.952	0.455	1.	2.957	5.048
3	Hrs.	2	3	1.952	0.455	1.	1.957	4.048
4	Hrs.	1	2	0.952	0.455	1.	0.957	3.048
5	Min.	10	60	7.602	22.756	50.	7.831	112.398
6	Min.	1	10	0.568	4.096	9.	0.610	19.432
7	Sec.	25	60	23.322	15.929	35.	23.482	96.678
8	Sec.	11	25	10.329	6.372	14.	10.393	39.671
9	Sec.	5	11	4.712	2.731	6.	4.740	17.288
A	Sec.	1	5	0.808	1.820	4.	0.826	9.192
B	Sec.	0.046	1	0.000	0.434	0.954	0.004	2.000

Summary

The simple one-parameter timeliness model does not represent the decay of value well. The Weibull seems to not represent loss well either. Of the three, the two-parameter exponential represents probable loss best. For this reason, the two-parameter exponential timeliness model is used in this paper. Table A.1 lists reference times for $\mathcal{T}_2(t)$ and all twelve $\mathcal{P}$ codes.

APPENDIX B. AUTOCORRELATION

The output of a queueing system simulation is similar to that of any other stochastic experiment except that it is autocorrelated. That is, each outcome is strongly influenced by its predecessors. This complicates analysis, since it increases both the variance and bias of most estimators and reduces the independence of independently-seeded replicates. These complications tend to lead to much longer simulation runs than if autocorrelation were not present. In addition, it frustrates efforts to estimate the shape of F(t). Notation introduced in this discussion is summarized in Table B.1

Selected Properties of Simple Queueing Systems

Before discussion autocorrelation, first consider the following properties of a single-server queue in which arrivals follow a Poisson distribution and service times follow some independent general distribution. Such systems are denoted (M/G/1) systems in Kendall notation. In spite of their simplicity, they serve to illustrate numerous features of C^3I system. All examples in this chapter are variations of the (M/G/1) queue.

Table B.1: Autocorrelation Notation

Term		Notation
$\{Y_i;\ i = 1, \ldots, n\}$	-	An output sequence from a simulation.
i_0	-	The point, beyond which the sequence may be considered covariance-stationary
$r(k)$	-	The lag correlation of Y_i and Y_{i+k} for a covariance-stationary process. See (B.9).
$\sigma^2_{Y_i}$	-	The variance of Y_i, $i = i_0, \ldots, n$.
$\mathcal{L}$	-	The critical lag, equal to the sum of all correlation coefficients, r(k). See (B.13).
$\hat{r}(k)$	-	The sample estimate of $r(k)$. See (B.14)
$\{X_j;\ j = 1, \ldots, N\}$	-	A sequence of batch means derived from $\{Y_i\}$.
q	-	The Von Neumann statistic. See (B.16).
$\{Z_i;\ i = 1, \ldots, n\}$	-	A sequence of cumulative means derived from $\{Y_i\}$.
$S_n(k)$	-	The difference between the n-th and k-th cumulative means.
$T_n(t)$	-	A standardized sequence, derived from $S_n(k)$. See (B.21).
$\hat{h}$	-	Schruben's test statistic. See (B.22).
I_0	-	The point beyond which the sequence $\{Y_i\}$ is essentially covariance-stationary and free of the effects of the initial conditions.

Distribution of residence time

The residence time T is the time a customer spends waiting for service, T_q, plus the time it spends in service, T_s. The Laplace-Stieltjes Transform (LST) maps a probability distribution (in t) to the complex plane (in s) with the function $\mathrm{E}[e^{-sT}]$. The LST of the residence time in a (M/G/1) queueing system is [46, §5.1.6]:

$$W^*[s;\rho,\lambda,b(t)] = \frac{(1-\rho)sB^*(s)}{s-\lambda[1-B^*(s)]} \tag{B.1}$$

where $B^*(s)$ is the LST of $b(t)$, the service time distribution. For example, if the service time follows an Erlang distribution with k phases and overall mean $\frac{1}{\mu}$, then since $B^*(s;k,\mu) = \left(\frac{1}{1+\frac{s}{k\mu}}\right)$ and $\lambda = \mu\rho$,

$$\begin{aligned} W^*(s;\rho,k,\mu) &= \frac{(1-\rho)s}{s\left(1+\frac{s}{k\mu}\right)^k - \mu\rho\left(1+\frac{s}{k\mu}\right)^k + \mu\rho} \\ &= \frac{1-\rho}{\sum_{n=0}^{k} a(n;k,\mu,\rho)s^n} \end{aligned} \tag{B.2}$$

with

$$a(n;k,\mu,\rho) = \left[\binom{k}{n} - \binom{k}{n+1}\frac{\rho}{k}\right]\left(\frac{1}{k\mu}\right)^n \tag{B.3}$$

noting that $\binom{k}{k+1} = 0$.

Moments of residence time

In simple cases, such as when service time is exponential, one may easily invert (B.1) to obtain the distribution of residence times. In other cases, expansion of e^{-sT} shows that

$$\mathrm{E}\left[T^n\right] = (-1)^n \left.\frac{d^{(n)}W^*(s)}{ds^n}\right|_{s=0} \tag{B.4}$$

Once again, if service time is Erlang-k, then

$$\mathrm{E}[T|k,\mu,\rho] = \frac{a(1;k,\mu,\rho)}{a(0;k,\mu,\rho)} \tag{B.5}$$

$$= \frac{k(2-\rho)+\rho}{2k\mu(1-\rho)}$$

$$\mathrm{E}[T^2|k,\mu,\rho] = \frac{2\left[a(1;k,\mu,\rho)^2 - a(0;k,\mu,\rho)a(2;k,\mu,\rho)\right]}{a(0;k,\mu,\rho)^2}. \tag{B.6}$$

Thus, the coefficient of variation

$$\varphi(\rho,k) = \sqrt{1-\frac{2a(0)a(2)}{a(1)^2}}$$

$$= \sqrt{1-\frac{4(k-1)(1-\rho)\left[3k-(k-2)\rho\right]}{3\left[2k-(k-1)\rho\right]^2}} \tag{B.7}$$

is clearly less than one, increases with ρ, decreases with k, and is independent of μ. Figure 3.18 Page 119 shows φ as a function of ρ for selected k. Note that the coefficient of variation is low only if the service time distribution has low variance and the offered load is also light.

The Nature of Autocorrelation

Let $\{Y_i, i = 1,\ldots,n\}$ be an output sequence from a simulation. The autocorrelation is a measure of the degree to which a given Y_i is similar to its neighbors. Let

$$\mathrm{Corr}\left[Y_i, Y_{i+k}\right] = \frac{\mathrm{Cov}\left[Y_i, Y_{i+k}\right]}{\sqrt{\mathrm{Var}[Y_i]\mathrm{Var}[Y_{i+k}]}}. \tag{B.8}$$

If there exists some i_0, such that the autocorrelation is independent of i for $i \geq i_0$, the process is said to be *covariance stationary* beyond i_0. For such sequences, the

lag autocorrelation function is defined as:

$$r(k) = \text{Corr}[Y_i, Y_{i+k}] \qquad i = i_0, \ldots, n-k \tag{B.9}$$

From its definition, $r(0) = 1$, $r(-k) = r(k)$ and $-1 \leq r(k) \leq 1$. Also, if a process is covariance stationary, then since $\sigma^2_{Y_i} = \text{Var}[Y_i] = \text{Cov}[Y_i, Y_i]$, $\sigma^2_{Y_i}$ is independent of i and

$$\begin{aligned} \text{Cov}\left[Y_i, Y_{i+k}\right] &= \frac{\text{Cov}[Y_i, Y_{i+k}]}{\sigma^2_Y}\ \sigma^2_Y \\ &= r(k)\,\sigma^2_Y \end{aligned} \tag{B.10}$$

where $\sigma^2_Y = \text{Var}(Y_i)$; $i = i_0, \ldots, n$.

The properties of $r(k)$ in queueing systems are discussed in [51, 52, 53, 54]. Among other results, these show that in stable queueing systems $r(k)$ is positive, a monotonically decreasing function of k and $\lim_{|k|\to\infty} r(k) = 0$.

The Effects of Autocorrelation

Inflation of estimator variance

If one estimates a parameter by the sample mean ($\overline{Y}$) and its generating process is covariance stationary, then following the argument in [55, p. 304], the variance of the estimate is

$$\begin{aligned} \text{Var}\left(\overline{Y}\right) &= \frac{1}{N^2}\,\text{Var}\left(\sum_{i=1}^{N} Y_i\right) \\ &= \frac{1}{N^2}\left[\sum_{i=1}^{N} \text{Var}(Y_i) + 2\sum_{k=1}^{N-1}\sum_{j=k+1}^{N} \text{Cov}(Y_k, Y_j)\right] \end{aligned}$$

$$= \frac{1}{N^2}\left[N\sigma_Y^2 + 2\sum_{k=1}^{N-1}(N-k)\sigma_Y^2 r(k)\right]$$

$$= \frac{\sigma_Y^2}{N}\left[\sum_{k=-(N-1)}^{N-1}\frac{N-|k|}{N}\,r(k)\right] \qquad \text{(B.11)}$$

If N is sufficiently large, $\lim_{k\to\infty} r(k) = 0$ implies that:

$$\text{Var}\left(\overline{Y}\right) \approx \frac{\sigma_Y^2}{N}\left[\sum_{k=-\infty}^{\infty} r(k)\right] \qquad \text{(B.12)}$$

providing, of course, the infinite sum converges. Denote the *critical lag*

$$\mathcal{L} \triangleq \sum_{k=-\infty}^{\infty} r(k) \qquad \text{(B.13)}$$

It is clear from (B.12) that $\mathcal{L}$ represents the average number of Y_i that are needed to obtain the equivalent of a single independent observation.

Initial bias

Each simulation run must begin in a well-defined state. The autocorrelation will cause the effect of that state to persist for some time. If the initial state is one of low probability, then the initial values of Y_i will be biased. Furthermore, since the system is not in steady-state, $\text{Corr}(Y_i, Y_{i+k})$ is not independent of i and (B.11) is not valid. This is a concern in non-terminating simulation.

Reduction of independence between replicates

If independently-seeded replicates start in a common state, the autocorrelation will cause the $Y_{m,i}$ of different replicates to be correlated. This is also a concern in non-terminating simulation.

Estimation of Autocorrelation and its Effects

Sample estimates of critical lag

Because of the central role of $r(k)$ in queueing system analysis, numerous methods have been presented to estimate its extent. For example, one can use the sample autocorrelation function

$$\hat{r}(k) \triangleq \frac{\hat{C}(k)}{\hat{C}(0)} \tag{B.14}$$

where

$$\hat{C}(k) \triangleq \sum_{i=1}^{N-k} \frac{\left(Y_i - \overline{Y}\right)\left(Y_{i+k} - \overline{Y}\right)}{N} \tag{B.15}$$

to estimate $r(k)$ although it tends to severely underestimate $r(k)$ if N is too small [56, p. 241ff]. One can then establish a trial value of $\mathcal{L}$ to group the Y_i into $N = \left[\frac{n}{\mathcal{L}}\right]$ batches with mean values $X_1, X_2, \ldots, X_N$. If the X_j are in fact uncorrelated, the expected value of the Von Neumann statistic

$$q\left(\{X_j\}\right) = \frac{\sum_{j=1}^{N-1}\left(X_j - X_{j+1}\right)^2}{\sum_{j=1}^{N}\left(X_j - \overline{X}\right)^2} \tag{B.16}$$

will be equal to 2. The hypotheses that $\mathrm{Corr}(X_j, X_{j+1}) = 0$ is rejected if $q(\{x_j\})$ is smaller than

$$q^*_{N,\alpha} = 2 - Z_\alpha\sqrt{\frac{4(N-2)}{(N-1)(N+1)}} \tag{B.17}$$

where Z_α is the upper α probability point of the standard normal distribution. Fishman [57] has reported that this test seems valid for N as small as 8, but noted

problems in coverage when traffic intensity is high. For this reason, Kleijnen [27, p. 67] recommends that N be at least 100.

Queueing theoretic estimates of critical lag

Pagurek and Woodside presented a method to compute $\mathcal{L}$ in a wide variety of (GI/G/1) queueing systems [54]. Using their reasoning, one can show that in (M/G/1) systems,

$$\mathcal{L} = 1 + \frac{\lambda\left(\mathrm{E}[T^3] - \mathrm{E}[T]\mathrm{E}[T^2]\right)}{(1-\rho)\mathrm{Var}(T)} \tag{B.18}$$

The analyst can use this result together with an approximate queueing characterization of the path(s) under consideration to estimate the mean value of $\mathcal{L}$.

For example, consider the (M/D/1) queue, in which service time is fixed. From (B.6) and (B.6), $\lim_{k\to\infty} \mathrm{E}[T] = \frac{2-\rho}{2\mu(1-\rho)}$ and $\lim_{k\to\infty} \mathrm{E}[T^2] = \frac{6-4\rho+\rho^2}{6\mu^2(1-\rho)^2}$. Thus, $\lim_{k\to\infty} \mathrm{Var}(T) = \frac{4\rho-\rho^2}{12\mu^2(1-\rho)^2}$. From (B.4), $\lim_{k\to\infty} \mathrm{E}[T^3] = \frac{4-\rho}{4\mu^3(1-\rho)^3}$. Thus, for the (M/D/1) model,

$$\mathcal{L} = \frac{4+2\rho}{(4-\rho)(1-\rho)^2} \tag{B.19}$$

Figure 3.18 on Page 119 illustrates $\mathrm{E}[T]$, φ, and $\mathcal{L}$ as functions of ρ with $\mu = 1$ in the (M/D/1) queue. Note that even with T_s fixed, $\mathcal{L}$ rises very quickly with ρ and affects output long before the system becomes congested.

Sample estimate of the extent of initial bias

There is no shortage of methods to detect the initial transient. Unfortunately, most of them do not work [39, 40, 41]. Schruben, however, has presented a promising

approach together with some validation in [58]. Basically, the method is to estimate I_0 by examining a visual trace of the composite of several replicates and then test the estimate for sufficiency.

Define Z_i to be the cumulative mean of the first i observations. Schruben models the Z_i as the sum of a signal due to a shift in the mean plus random noise. The possible signal is detected by transforming the Z_i to a standardized Brownian bridge. The series

$$S_n(k) \triangleq \begin{cases} Z_n - Z_k & \text{if } 0 < k \leq N \\ 0 & \text{if } k = 0 \end{cases} \tag{B.20}$$

is an estimate of the mean cumulative bias after considering $Y_1, \ldots, Y_k$ from a sample of size n. Define the critical value of t to be the smallest one that maximizes the standardized test sequence

$$T_n(t;\sigma) \triangleq \frac{(tn)S_n(tn)}{\sigma\sqrt{n}}. \tag{B.21}$$

Here $t = \frac{k}{n}$ is standardized so that $0 \leq t \leq 1$ and σ is the variance of the random noise. Since n and σ are fixed in value, the maximum of $T_n(t;\sigma)$ occurs at the largest value of $kS_n(k)$.

Define, $\hat{k}$ to be the smallest value of $k = 1, 2, \ldots, n$ that maximizes $kS_n(k)$ and let $\hat{s} = \frac{\hat{k}S_n(\hat{k})}{\sqrt{n}}$. [58, §3.1] shows that $3\hat{h}\frac{\hat{\sigma}^2}{\sigma^2}$ has an approximate χ^2_3 distribution when no initialization bias is present, where

$$\begin{aligned} \hat{h} &\triangleq \frac{\hat{s}^2}{3\hat{\sigma}^2\hat{t}(1-\hat{t})} \\ &= \frac{n^2\hat{s}^2}{3\hat{\sigma}^2\hat{k}(1-\hat{k})} \end{aligned}$$

$$= \frac{n\hat{k}S_n^2(\hat{k})}{3\hat{\sigma}^2(1-\hat{k})}. \tag{B.22}$$

The analyst may use this to form one- or two-sided tests for bias.

Computing $\hat{h}$ requires an estimate of σ. Because of this, Schruben also presents a simpler ratio test. The h-ratio test computes $\hat{h}_f$ from the first half of the output and $\hat{h}_s$ from the second. Under the hypothesis that σ^2 is the same in both sub-samples, the ratio $\frac{\hat{h}_f}{\hat{h}_s}$ has an approximate F-distribution with 3 and 3 degrees of freedom.

To assure proper coverage, Schruben recommends that one compute $\hat{h}_f$ and $\hat{h}_s$ from two independent samples. To achieve sufficient power, he recommends that each sample be a composite of at least five independently-seeded runs. Finally, to reduce the effect of noise, he recommends that the $Y_{m,i}$ be serially grouped into batches of about five prior to computing the Z_j.

Estimates of loss of independence

At present, there does not seem to be a practical test for initial correlation among the independently-seeded replicates. Analyst could use the Von Neumann statistic, for example, but they would need about 100 replicates. Usually, the analyst simply examines smoothed traces of system performance and revises I_0 upward correlation between replicates seems to persist beyond the region of initial bias.

APPENDIX C. SUPPLEMENTAL DETAILS OF EXAMPLE 2.1

The following describes details of Example 2.1 that were not essential to the main discussion. The methods follow those described in [14], except where they are modified or augmented by the proposed method.

Modeling Phase

Problem Translation

Conversions This section derives the values in Table C.1 A crucial part of the proposed method is that all C^2 data is stated as intervals with some presumed subjective probability. At this time, all CDB data is given as point values with no measure of precision. For the sake of discussion, assume all true values are within ±20% of the stated value with 95% subjective probability. The translation of call frequency in calls per day to mean interarrival time in seconds is straightforward. The range is derived from the limits of the arrival rate. The method to convert service times depends upon the type of call. Data messages would be transmitted in packets. Assuming there is a twenty percent overhead for the packet and signaling, the mean service time per message in seconds would be:

$$\theta_S = m \text{ kBytes} \times \frac{8000 \text{ bits}}{\text{kByte}} \times 1.2 \times \frac{1 \text{ second}}{9600 \text{ bits}}$$

$$= \; m \text{ seconds} \tag{C.1}$$

where m is the message length in kBytes, stated in Table 2.2. In the case of voice traffic, [14, Appendix B, p. 8] states that one minute of conversation is equivalent to one thousand bytes. Hence, the service time for voice messages in seconds is just $60m$.

Caller behavior Time between recalls and deadlines are not addressed in the CDB. Because data traffic simply queues, the mean time between recalls, θ_r, affects only voice traffic. Table C.2 lists assumed ratios between t_ℓ and the time between recalls. Let the mean time between recalls $\theta_r = t_\ell F_{\mathrm{COF}}$. Because θ_r depends upon two CDB parameters, its subjective probability is 0.90. Deadlines affect all traffic. Let t_d be the time at which $\mathcal{V}(t) = 0.05$. In the case of the $\mathcal{T}_2(t)$ model,

$$t_d \; = \; t_0 - \tilde{t} \ln\left(\frac{0.05}{\mathcal{I}}\right). \tag{C.2}$$

Uncertainty An important element of the method is the assumption that all C^2 data is stated as intervals. At this time, there is no presumed precision for CDB parameters. For the sake of this example, assume experts have agreed on intervals for all values are ±20% with 0.95 subjective probability.

If the arrival rates and call lengths vary ±20%, the interarrival and service rates will also vary about ±20%. However, assuming the recall factor and $\mathcal{T}_\ell$ vary ±20% implies θ_r will vary from −36% to +44%. Finally, if only t_ℓ and t_h vary ±20%, so will t_d. However, since $\mathcal{I}$ can also vary, the range of t_d varies as shown in Table C.1. Finally, assume distributions are gamma with $0.5 < \varphi < 1.5$.

Table C.1: Traffic Specifications, in Seconds, for Example 2.1

Need Line (i)	Mean Time Between Calls[a b] (θ_{a_i})	Mean Message Service Time[a b] (θ_{s_i})	Mean Time Between Recalls[b c] (θ_{r_i})	Deadline (t_{d_i})		
				Low	Nominal	High
1	3600	480	300	4400	6440	10000
2	21600	4320	3600	10460	14200	19770
3	90	20	–	696	940	1333
4	360	75	–	4080	5500	7730
5	21600	1920	1800	7580	10600	15450
6	90	32	–	696	940	1333
7	180	16	–	616	770	924
8	21600	2880	3600	10450	14200	19770
9	18	4	–	26	34	46
10	36	8	–	62	82	112
11	3600	24	300	4400	6440	10000
12	1800	120	300	4400	6440	10000
13	60	4	–	753	1110	1742
14	120	12	–	4080	5500	7730
15	1800	96	30	753	1110	1742
16	72	4	–	696	940	1333
17	180	8	–	4080	5500	7730
18	7200	120	3600	7200	9500	12730
19	180	240	6	27	38	56
20	9	4	–	12	17	25

[a] Accurate to ±20% with subjective probability 0.95

[b] distribution is gamma with $\varphi \in [0.5, 2]$ and nominal value 1.0

[c] Accurate −36% to +44% with subjective probability 0.90

Table C.2: Cost of Failure Parameters

Cost of Failure	Recall Factor (F_{COF})			Importance Weight ($\mathcal{I}$)		
	Low	Nominal	High	Low	Nominal	High
I	0.40	0.50	0.60	1.00	1.00	1.00
C	0.80	1.00	1.20	0.25	0.50	0.70
E	1.80	2.00	2.40	0.06	0.25	0.50

Bounding system parameters

Task completion time is the sum of two random variables: T_s, the service time and T_q, the queueing time. The service time is a characteristic of the needline and is not affected by the load. The queueing time, however, is a function of the path the needline uses and the total load offered to each node along that path. Although an individual needline's queueing time may be affected by its recall rate and deadline, it is more strongly affected by the combined set of call initiations and service times for all needlines using its shared resources. Three elements affect each needline's contribution to the load. The first is the service time distribution, the second the queueing time distribution, and the third is its timeliness function, which affects its recall rate and deadline. One would expect that the contribution of a needline to the load would depend upon which resources it requires and be in proportion to its relative arrival rate and mean service time.

Table C.3 lists the utilization factor $\left(\rho_i\right)$ for each needline. Utilization is the ratio of the mean service time to the mean time between calls. Although this figure ignores the effects of queueing and recalls, it is a measure of the relative impact of each needline and how busy each resource can get. Table C.3 also lists relative utilization, a measure of how much a needline affects performance within its class.

Table C.4 shows the effects of these needlines on the key resources, varying the input and service rates by ±20%.

First Screening

Experiment design

The parameter array The parameter, or inner, array represents possible network models. M represents the call-back method, with −1 representing the random call-back option and +1, automatic call-back. V represents the number of voice channels and D the number of data channels. Since this is a screening stage, the last two factors could be at two levels. However, a minimal design in which the $M \times V$ interaction would be distinct requires eight runs. The modified L_9 array from [59], and techniques from [38] and [27] in Table C.5, allows tests of the latter two factors at three levels while leaving all two-factor interactions distinct.

Considering the maximum load in Table C.4, the least reasonable value of V and D is three. The other values were six and nine. The design will show the main effects of the call-back method and number of each kind of trunk as well as the interaction between method and number of trunks. Although M would not be expected to interact with D nor V with D, as shown in Table C.5, one can test for these interactions.

The noise array The noise, or outer, array represents variations of the offered load in this experiment. At this stage, the analyst one can consider each needline at the two extremes of offered load. All load parameters associated with each needline will be considered as a single factor with two levels, corresponding to

Table C.3: Absolute and Relative Utilization Factors

Need Line	Utilization Factor (ρ)							
	Absolute				Relative			
	All Calls	Trunk Calls			All Calls	Trunk Calls		
		Total	Voice	Data		Total	Voice	Data
1	0.133	—	—	—	0.033	—	—	—
2	0.200	—	—	—	0.049	—	—	—
3	0.222	—	—	—	0.055	—	—	—
4	0.208	—	—	—	0.051	—	—	—
5	0.089	0.089	0.089	—	0.022	0.030	0.058	—
6	0.356	0.356	—	0.356	0.088	0.122	—	0.258
7	0.089	0.089	—	0.089	0.022	0.030	—	0.065
8	0.133	—	—	—	0.033	—	—	—
9	0.222	—	—	—	0.055	—	—	—
10	0.222	0.222	—	0.222	0.055	0.076	—	0.161
11	0.007	—	—	—	0.002	—	—	—
12	0.067	0.067	0.067	—	0.016	0.023	0.043	—
13	0.067	0.067	—	0.067	0.016	0.023	—	0.048
14	0.100	0.100	—	0.100	0.025	0.034	—	0.073
15	0.053	0.053	0.053	—	0.013	0.018	0.035	—
16	0.056	0.056	—	0.056	0.014	0.019	—	0.040
17	0.044	0.044	—	0.044	0.011	0.015	—	0.032
18	0.017	—	—	—	0.004	—	—	—
19	1.333	1.333	1.333	—	0.328	0.457	0.865	—
20	0.444	0.444	—	0.444	0.109	0.152	—	0.323
Total	4.063	2.920	1.542	1.378	1.000	1.000	1.000	1.000

Table C.4: Lower Bounds on Resource Utilization Under Various Loads in Example 2.1

Resource	Capacity	Offered Load and Utilization					
		Low Limit		Nominal Value		High Limit	
		Load	ρ	Load	ρ	Load	ρ
Node A	10	1.56	0.156	2.34	0.234	3.51	0.351
Node B	5	0.74	0.148	1.11	0.222	1.67	0.334
Node C	5	0.88	0.176	1.32	0.268	1.98	0.396
Node X	20	1.43	0.072	2.15	0.108	3.23	0.162
Node Y	20	1.33	0.067	2.00	0.100	3.00	0.150
Voice Trunk	3	1.03	0.343	1.54	0.513	2.31	0.770
	6	1.03	0.172	1.54	0.257	2.31	0.385
	9	1.03	0.121	1.54	0.171	2.31	0.257
Data Trunk	3	1.19	0.397	1.78	0.593	2.67	0.890
	6	1.19	0.198	1.78	0.312	2.67	0.445
	9	1.19	0.132	1.78	0.198	2.67	0.297

Table C.5: The Inner (Parameter) Array for the First Screening Stage

Point	Treatment	Factor Levels			Interactions		
		M	V	D	$M \times V$	$M \times D$	$D \times V$
1	(l)	−	−	−	+	+	+
2	Md	+	−	0	−	0	0
3	D	−	−	+	+	−	−
4	Mv	+	0	−	0	−	0
5	vd	−	0	0	0	0	0
6	vD	−	0	1	0	−	0
7	V	−	+	−	−	+	−
8	Vd	−	+	0	−	0	0
9	MVD	+	+	+	+	+	+

Table C.6: Factor Screening Levels

Parameter	Least Load	Greatest Load	Meaning
θ_a	$1.2\theta_{a_0}$	$0.8\theta_{a_0}$	Mean time between arrivals.
θ_s	$0.8\theta_{s_0}$	$1.2\theta_{s_0}$	Mean service time.
θ_r	$1.2\theta_{r_0}$	$0.8\theta_{r_0}$	Mean time between recalls.
t_d	$0.8t_{d_0}$	$1.2t_{d_0}$	Recall deadline.
ϕ_a	0.5	2.0	Coefficient of variation for the arrival distribution.
ϕ_s	2.0	0.5	Coefficient of variation for the service distribution.

the lightest and the heaviest probable load. The levels associated with these extremes are shown in Table C.6. Note that instead of examining extremes of θ_r and t_d as consequences of $\mathcal{P}$ and $\mathcal{I}_0$, the design simply considers variations of $\pm 20\%$ in their values.

To further simplify the design, the analyst grouped the needlines by call type. L_{D} represents the relative level for all data calls, and L_{V} that of voice calls. A full-factorial design with two groups requires only four points. Yet, with this noise array, the analyst can test the assumption that data and voice traffic do not interact.

Simulation results

Each of the 36 sample points was simulated for five independently-seeded runs of about 24 simulated hours each. Five replicates of each point were needed in order to have sufficient power for Schruben's F-ratio test. In no case did I_1 exceed 1% of a run. The 24-hour period was necessary to give the infrequent voice traffic a chance to enter into the simulation. In most cases, the 24-hour runs were sufficient to determine L_1.

However, in eight cases, the runs were lengthened to no more than 30 simulated hours to provide enough observations for Von Neumann's test of independence. Finally, in each replicate, the first 1% of the observations were discarded and the remainder batched to produce twenty observations per replicate. These were analyzed through SAS.

The variation within each replicate was much less than that between the replicates. The cause of this is uncertain. One possibility is that the infrequent voice calls introduced a correlation that was masked by the data calls. Another is that the Von-Neumann statistic has insufficient power, even with 100 points. A third is that "overbatching" may have reduced the sample variance. Regardless, a conservative approach was taken, treating the average of the twenty observations in each run as a single observation.

The obvious question was, would this reduction in the number of observations significantly reduce the power of the ANOVA? The Type II error of ANOVA can be determined from [60] as follows. First, the non-central F distribution depends upon ν_2, the denominator d.f., and ϕ, Tang's non-centrality parameter. Because the distribution is essentially unchanged for any $\nu_2 > 90$, the drop from 884 to 164 d.f. had no great affect on tables values. Next, for a main effect, H_0: $\beta_i = 0$, $i = 1, \ldots, \ell$,

$$\phi(\beta; N, \ell, \sigma) = \sqrt{\frac{N}{\ell^2} \sum_{i=1}^{\ell} \frac{\beta_i^2}{\sigma^2}} \tag{C.3}$$

where ℓ is the number of levels of factor β and N is the total number of observations. For a crossed effect, H_0: $\gamma_{i,j} = 0$, $i = 1, \ldots, \ell$, $j = 1, \ldots, m$,

$$\phi(\gamma; N, \ell, m, \sigma) = \sqrt{\frac{N}{\ell m \left[(\ell - 1)(m - 1) + 1\right]} \sum_{i=1}^{\ell} \sum_{j=1}^{m} \frac{\gamma_{i,j}^2}{\sigma^2}}. \tag{C.4}$$

Table C.7: Analysis of Variance of the Screening Results for Example 2.1

Source	df	Type III Sum of Squares	Mean Square	F Value	Pr > F
Model	15	0.074946	0.00499	55.51	0.0001
Error	164	0.014761	0.00009		
Corrected Total	179	0.089707			
M	1	0.003746	0.003746	41.62	0.0001
D	1	0.000486	0.000486	5.40	0.0213
$D \times M$	1	0.000023	0.000023	0.25	0.6145
V	1	0.001468	0.001468	16.31	0.0001
$V \times M$	1	0.006330	0.006330	70.32	0.0001
$D \times V$	1	0.001327	0.001327	14.75	0.0002
L_V	1	0.000026	0.000026	0.29	0.5886
$M \times L_V$	1	0.000040	0.000040	0.44	0.5077
$D \times L_V$	1	0.000070	0.000070	0.78	0.3792
$V \times L_V$	1	0.000233	0.000233	2.59	0.1095
L_D	1	0.013139	0.013139	145.97	0.0001
$M \times L_D$	1	0.000006	0.000006	0.06	0.8049
$D \times L_D$	1	0.011963	0.011963	132.91	0.0001
$V \times L_D$	1	0.000037	0.000037	0.41	0.5228
$L_V \times L_D$	1	0.000067	0.000067	0.75	0.3891

Let the sum in either case equal unity. That means the region of acceptance for the Type II error is a hypersphere of radius σ. In this case, the Type II error for all single factors and most crossed factors is essentially zero. That for the $D \times V$ interaction, however, is about 19%. Since this effect is significant, the Type II error is very small, in spite of the loss of 720 degrees of freedom.

Table C.8: Sensitivity of Mean Timeliness at Nominal Values of θ_a and $\mathcal{I}$ in Example 2.1

Needline	On Task?	Call Type	COF Code	Sensitivity
1	Yes	Voice	I	0.001
2	Yes	Voice	I	0.000
3	Yes	Data	C	0.027
4	No	Data	C	0.007
5	Yes	Voice	I	0.000
6	Yes	Data	C	0.027
7	No	Data	E	0.007
8	Yes	Voice	I	0.000
9	No	Data	C	0.136
10	No	Data	C	0.068
11	Yes	Voice	I	0.001
12	Yes	Voice	I	0.003
13	Yes	Data	I	0.081
14	No	Data	C	0.020
15	Yes	Voice	I	0.003
16	No	Data	C	0.034
17	Yes	Data	C	0.014
18	Yes	Voice	C	0.000
19	No	Voice	I	0.027
20	No	Data	I	0.543

Second Screening

Experiment design

Although the second screening still concentrates on estimating $E_S[\mathcal{T}(T)]$ with p_ℓ, it differs from the first in four ways:

1. The range of the number of trunks is smaller,
2. The range of the noise effects is reduced,
3. The sensitivity of $E_S[\mathcal{T}(T)]$ to $\mathcal{T}(T)$ is checked,
4. The simulation analysis is more efficient, and
5. The baseline model is examined.

These differences are explained below.

Number of trunks The first screening suggests that the greatest gain in performance occurs for $3 \leq D \leq 6$ and that D has a much greater effect on p_ℓ than V. This suggests that the eight-trunk option may be cost-effective. In the second screening, $C = 8$, $4 \leq D \leq 6$ and $V = 8 - D$. The design, shown in Table C.9, is a full-factorial balanced design that considers M, D, and $M \times D$. Also, this design has one-third fewer rows than the first screening. This causes Type II error for the $M \times D$, $D \times L_D$, and $D \times L_V$ interactions to increase to about 4.5%. The Type II error for the single factors and other interactions are still nearly zero.

Load range Assume that the analyst, troubled by the large effect of load changes on p_ℓ, asks C^2 experts to investigate the load levels for four high-traffic

Table C.9: The Inner (Parameter) Array for the Second Screening Stage

Point	Treatment	Factor Levels		Interaction
		M	D	$M \times D$
1	(l)	−	−	+
2	M	+	−	−
3	d	−	0	0
4	Md	+	0	0
5	D	−	+	−
6	MD	+	+	+

needlines. The results are shown in Table 2.6 on Page 62. Although there was no improvement for Needline 6, the other three account for nearly 70% of the total and data loads.

Sensitivity of $\mathrm{E_S}[\mathcal{T}(T)]$ to $\mathcal{T}(t)$ In the former screening, one could test the sensitivity of $\mathrm{E_S}[\mathcal{T}(T)]$ to changes in $\mathcal{I}$, but not the sensitivity to models of $\mathcal{T}(t)$ which change t_ℓ. In the second screening, the analyst assumes t_0 and $\tilde{t}$ can both vary ±20%, which means t_ℓ can vary ±20%. Each estimate of p_ℓ is measured at $0.8t_\ell$, t_ℓ, and $1.2t_\ell$. Note that since $C_1 = \left[t_0 + \tilde{t}(1 - V_\ell)\right] \left[t_0 - \tilde{t} \ln V_\ell\right]$, its value is the same for all three points

Efficiency Except when $D = 6$, the parameter points are within the extremes of the first model. Because the results of the first screening were so uniform in nature, each replicate was run for fifteen simulated minutes (about 1% of a 24-hour day), then after statistics were discarded, run for 24 additional simulated hours. The resulting replicate averages were comparable to the initial screening results.

Baseline Model It seems that there is slight improvement beyond six data trunks. If this is true, then the eight-trunk model is viable. To check this assumption, baseline runs were made, with scheduled recall and $V = D = 99$. Since there were never more than 50 calls in this model during simulation, these runs represent the very best possible trunk configuration. The $\mathrm{E_S}[\mathcal{T}(T)]$ for this model is the standard of performance.

Simulation results

The results were more mixed in the second screening than in the first. Table C.10 displays the chance probabilities for all six estimated quantities. The data load seems to have a great effect and may dominate the over-all results. If t_ℓ is reduced too much, its effect is also significant. Parameters and the voice load seem to have little consistent effect. On the other hand, the results imply the eight-channel trunk could be a cost-effective choice.

Figure 2.18 on Page 63 and Figure C.1 illustrate these results graphically. Figure 2.18 illustrates the baseline and model levels for $\mathrm{E_S}[\mathcal{T}(T)]^-$. Comparing these levels shows the eight-channel trunk delivers very high effectiveness. Figure 2.18 shows the data load affects $\mathrm{E_S}[\mathcal{T}(T)]^-$ more than the voice load, but much less than in the first screening.

Figures 2.19 and 2.20 on Pages 63 and 64 illustrate the effect of different timeliness and effectiveness models on $\mathrm{E}[\mathcal{V}(T)]^-$. Changing the importance ratio reduces the value of the estimate, but increasing t_ℓ simultaneously increases both $\mathrm{E_S}[\mathcal{T}(T)]^-$ and its variance. In either case, the effects of uncertainty in the effectiveness model are about the same as those due to sampling variance and offered load.

Table C.10: Analysis of Variance of the Second Screening Results

Source	df	Probability That Effects on $E_S[\mathcal{T}(T)]^-$ are Due to Chance					
		At $0.8 \times t_\ell$		At t_ℓ		At $1.2 \times t_\ell$	
		$\mathcal{I}$ =I	$\mathcal{I}$ =C	$\mathcal{I}$ =I	$\mathcal{I}$ =C	$\mathcal{I}$ =I	$\mathcal{I}$ =C
Model	10	0.0001*	0.0001*	0.0001*	0.3475	0.0001*	0.0001*
M	1	0.7491	0.6601	0.9128	0.7568	0.9509	0.7729
D	1	0.1901	0.2128	0.0110†	0.2206	0.0002*	0.0884
$D \times M$	1	0.4691	0.8507	0.5593	0.9879	0.5629	0.9504
L_V	1	0.0434†	0.5314	0.0152†	0.3670	0.0154†	0.3535
$M \times L_V$	1	0.7154	0.6392	0.7932	0.5414	0.9864	0.4884
$D \times L_V$	1	0.0289†	0.6683	0.0090‡	0.5373	0.0061‡	0.5033
L_D	1	0.0001*	0.0001*	0.0003*	0.8515	0.0071‡	0.0001*
$M \times L_D$	1	0.7411	0.6266	0.7546	0.7694	0.6642	0.5566
$D \times L_D$	1	0.0201†	0.8399	0.0437†	0.7873	0.0623	0.6404
$L_V \times L_D$	1	0.1153	0.4140	0.1430	0.1707	0.1352	0.0201†

Superscript indicates significance at † = 5%, ‡ = 1%, and * = 0.1%

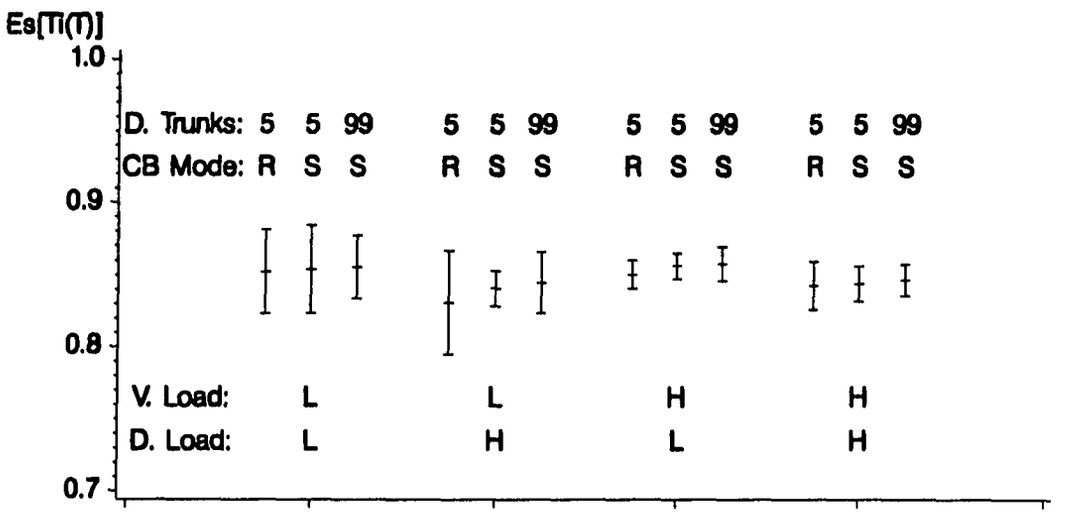


Figure C.1: 95% Confidence Intervals for Load Effects on $E_S[\mathcal{T}(T)]^-$ for Selected Models from the Second Screening

Final Analysis

The final phase is to compare the candidate systems. The figure of merit will be the amount of initial value that the system retains. This will be estimated through the two-parameter hybrid approximation $\mathrm{E_H}[\mathcal{V}(T)]$.

Procedure

The experimental plan and models were identical to that of the second screen, except that both p_ℓ and ζ were estimated for each of the twelve needlines that participated in the given task. Now

$$\mathrm{E}\left[\mathcal{V}\left(T;\mathcal{T}(t)\right)|\mathcal{I}_n,\lambda_n,T_n;\ n=1,\ldots,N\right] \;=\; \sum_{n=1}^{N}\mathcal{I}_n\left(\frac{\lambda_n}{\lambda}\right)\mathrm{E}\left[\mathcal{T}\left(T_n\right)\right] \quad \text{(C.5)}$$

where n, $\mathcal{I}_n$, λ_n, and T_n are the needline's index, importance weight, arrival rate, and system time, respectively. Furthermore, the summation is over those needlines involved in the task and $\lambda = \sum_{n=1}^{N}\lambda_n$. The next step was to estimate $\mathrm{E}[\mathcal{T}(T_n)]$ for each needline from the p_ℓ and ζ estimates and to estimate the weighted average of timeliness, $\mathrm{E}_i[\mathcal{T}(T)]$, as well as $\hat{w}_i = \frac{\widehat{\lambda_i}}{\lambda}$, within each importance group. Then

$$\hat{\mathrm{E}}\left[\mathcal{V}\left(T;\mathcal{T}(t)\right)\middle|\mathcal{I}_i,\hat{w}_i,T_i;\ i=1,2,3\right] \;=\; \sum_{i=1}^{3}\mathcal{I}_i\hat{w}_i\hat{\mathrm{E}}\left[\mathcal{T}(T_i)\right]. \quad \text{(C.6)}$$

This formulation allows the analyst to independently estimate the effect of changing $\mathcal{I}$ values. Finally, the portion of initial value retained was estimated by:

$$\hat{\mathrm{E}}_r\left[\mathcal{V}(T)\middle|\hat{\mathrm{E}}[\mathcal{V}(T)],\mathcal{I}_i,\hat{w}_i;\ i=1,2,3\right] \;=\; \frac{\hat{\mathrm{E}}[\mathcal{V}(T)]}{\sum_{i=1}^{3}\mathcal{I}_i\hat{w}_i}. \quad \text{(C.7)}$$

Table C.11: ANOVA Results for $\widehat{E_S}[\mathcal{T}(T)]$ at Nominal Importance Values

Source	df	Probability That Effects are Due to Chance		
		At $0.8t_0$ and $0.8\tilde{t}$	At t_0 and $\tilde{t}$	At $1.2t_0$ and $1.2\tilde{t}$
Model	10	0.0001*	0.0001*	0.0001*
M	1	0.9542	0.9699	0.8393
D	1	0.8064	0.1503	0.0152†
$D \times M$	1	0.5486	0.5985	0.4855
L_V	1	0.8194	0.9652	0.3597
$M \times L_V$	1	0.2170	0.1013	0.1877
$D \times L_V$	1	0.6456	0.7348	0.2194
L_D	1	0.0014‡	0.0001*	0.0001*
$M \times L_D$	1	0.8198	0.7090	0.7690
$D \times L_D$	1	0.2728	0.4523	0.5821
$L_V \times L_D$	1	0.5037	0.3919	0.6427

Superscript indicates significance at † = 5%, ‡ = 1%, and * = 0.1%

Results

Table C.11 lists ANOVA results for all three sets of t_0 and $\tilde{t}$ at nominal values of $\mathcal{I}_i$. Aside from the data load, relative effectiveness is not sensitive to model or load factors. Figure C.2 shows 95% C.I.s for $E_S[\mathcal{T}(T)]$ at nominal values of t_0, $\tilde{t}$, and $\mathcal{I}_i$. The response of each model option is indistinguishable from that of the others as well as that of the baseline model. Also, all C.I.s exclude $E_S[\mathcal{T}(T)] = 0.8$. Thus, for the nominal $\mathcal{T}(t)$ function and $\mathcal{I}_i$ values all six systems provide adequate effectiveness. Figures C.3 and C.4 show that at each load level, neither the recall mode nor the number of trunks affects the 95% C.I.s for $E_S[\mathcal{T}(T)]$ significantly.

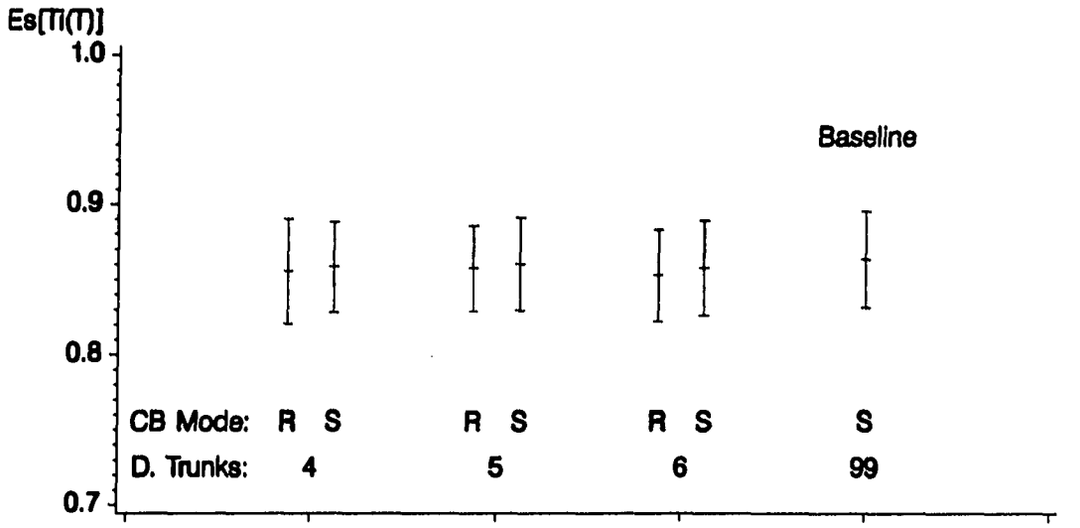


Figure C.2: 95% Confidence Intervals for the Model Effects on $\mathrm{E_S}[\mathcal{T}(T)]$ at Nominal Values of t_0, $\tilde{t}$, and $\mathcal{I}_i$

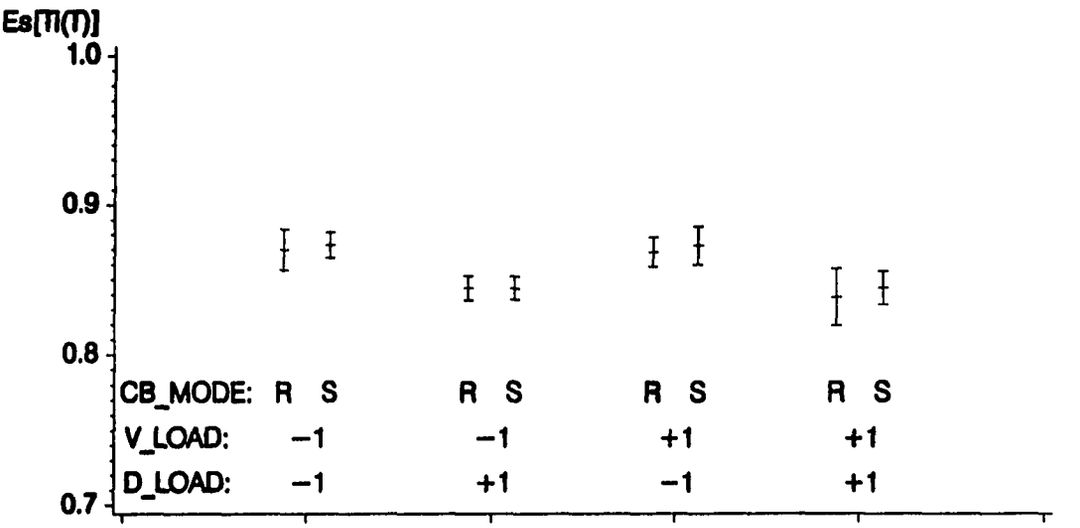


Figure C.3: 95% Confidence Intervals for the Offered Load and Recall Method Effects on $\mathrm{E_S}[\mathcal{T}(T)]$ at Nominal Values of t_0, $\tilde{t}$, and $\mathcal{I}_i$

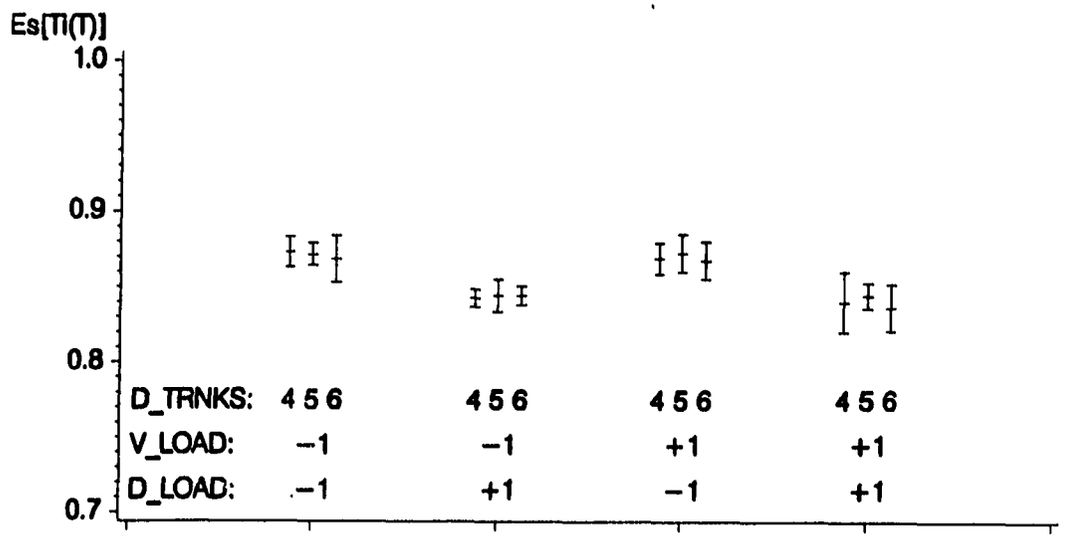
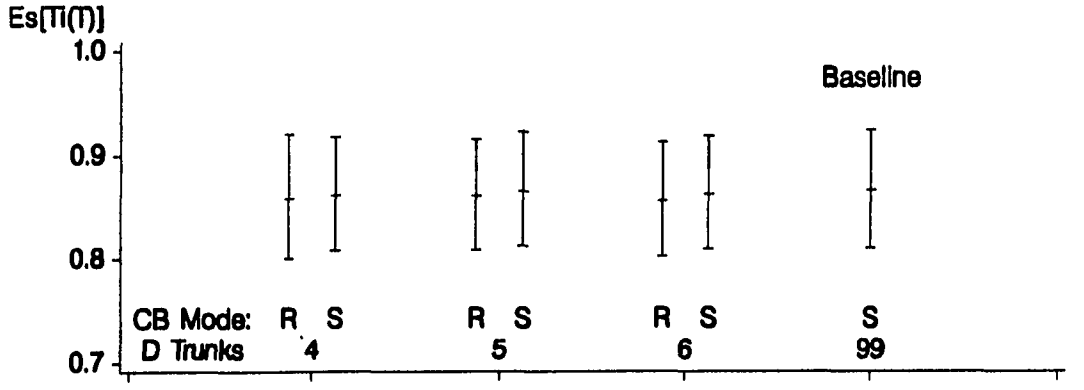


Figure C.4: 95% Confidence Intervals for the Offered Load and Number of Data Trunks Effects on $\mathrm{E_S}[\mathcal{T}(T)]$ at Nominal Values of t_0, $\tilde{t}$ and $\mathcal{I}_i$

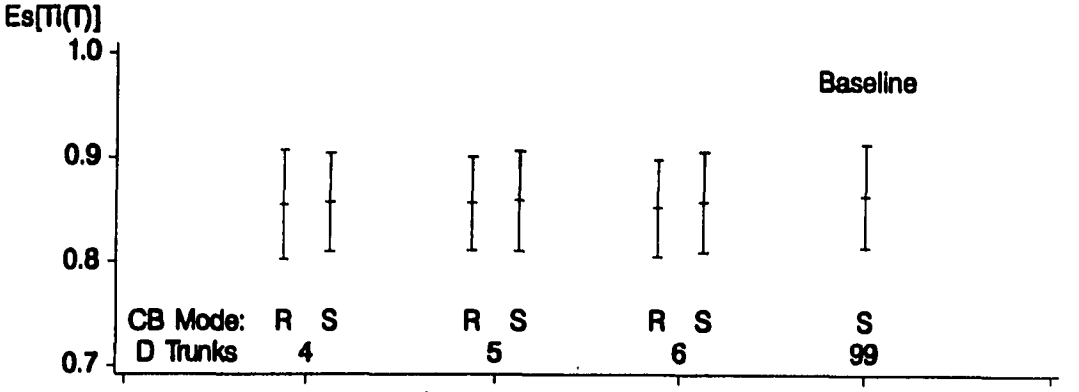
Table C.12: $\hat{\mathrm{E}}[\mathcal{V}(T)]$ at Extreme Importance Values

		Probability That Effects are Due to Chance					
		At $\mathcal{I}_\mathrm{I} = 2.0$ and $\mathcal{I}_\mathrm{C} = 1.5$			At $\mathcal{I}_\mathrm{I} = 16$ and $\mathcal{I}_\mathrm{C} = 4$		
Source	df	At $0.8t_0$ and $0.8\tilde{t}$	At t_0 and $\tilde{t}$	At $1.2t_0$ and $1.2\tilde{t}$	At $0.8t_0$ and $0.8\tilde{t}$	At t_0 and $\tilde{t}$	At $1.2t_0$ and $1.2\tilde{t}$
Model	10	0.0001*	0.0001*	0.0001*	0.0001*	0.0001*	0.0001*
M	1	0.8639	0.8111	0.9729	0.6944	0.6596	0.6650
D	1	0.8396	0.4286	0.0717	0.3438	0.0191†	0.0012‡
$D \times M$	1	0.7515	0.8538	0.6469	0.3246	0.3205	0.3164
L_V	1	0.8852	0.8130	0.3995	0.7631	0.6789	0.3455
$M \times L_\mathrm{V}$	1	0.2860	0.1384	0.2621†	0.1501	0.0731	0.1238
$D \times L_\mathrm{V}$	1	0.7538	0.9992	0.2776	0.5138	0.4193	0.1731
L_D	1	0.0042‡	0.0001*	0.0001*	0.0004*	0.0001*	0.0001*
$M \times L_\mathrm{D}$	1	0.9427	0.7683	0.7626	0.7174	0.6859	0.8231
$D \times L_\mathrm{D}$	1	0.3995	0.6145	0.7237	0.1465	0.2773	0.4190
$L_\mathrm{V} \times L_\mathrm{D}$	1	0.6794	0.0010*	0.7373	0.3247	0.2894	0.5113

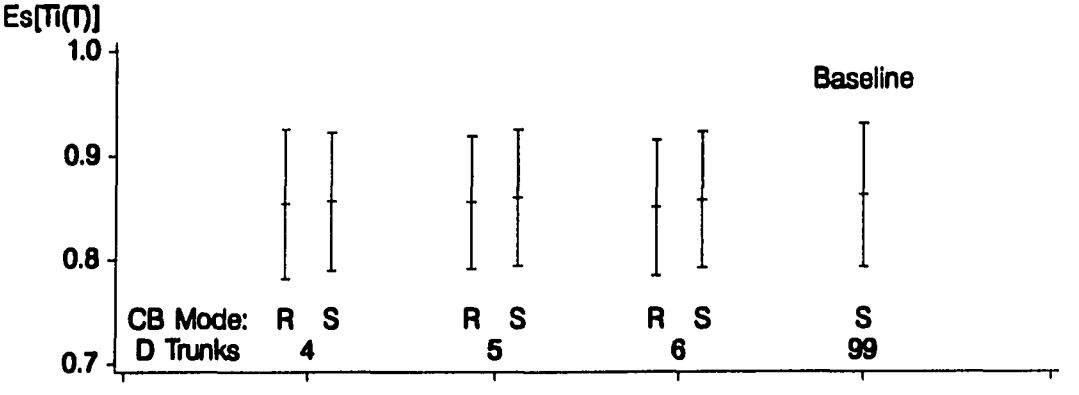
Superscript indicates significance at † = 5%, ‡ = 1%, and * = 0.1%



a. Over a Range of Timeliness Functions



b. Over a Range of Importance Ratios



c. Over a Range of Timeliness Functions and Importance Ratios

Figure C.5: Pseudo 95% Confidence Intervals for Model Effects on $E_S[\mathcal{T}(T)]$