AN ABSTRACT OF THE THESIS OF

William H. Stewart for the degree of Doctor of Philosophy

in Statistics presented on March 13, 1979

Title: A DISTRIBUTION-FREE APPROACH FOR GROUPED SURVIVAL DATA: ANALYSIS AND CALCULATION OF EFFICIENCY

Abstract approved: Redacted for privacy

Donald A. Pierce

A distribution-free analysis is proposed for inference concerning treatment effects in factorial survival experiments in which the recorded data are grouped by time intervals. A grouped data model is built by applying the continuous-time Cox regression and life model. This models the treatments to have multiplicative effects, possibly time dependent, on some unrestricted, hence distribution-free, underlying hazard function. By making certain good approximations a likelihood expression is obtained which depends only on parameters associated with the treatment effects. This allows large sample maximum likelihood inference about the treatment effects to proceed unburdened by nuisance parameters associated with the underlying hazard function. It is shown that, despite the approximation, the estimators obtained are consistent. The main error caused by the approximation concerns variance estimates, but this error will be small for most practical problems.

When the underlying hazard function can be smoothly modeled by some parametric form, the efficiency of the distribution-free analysis is found to be very high. This efficiency is investigated by considering smoothing restrictions on the previously unrestricted parameters associated with the underlying hazard function of the grouped data model. It

is shown that the distribution-free analysis is geared to give high efficiency for local alternatives to the hypothesis of no treatment effect and for smooth modeling allowing monotone decreasing hazard functions.

A DISTRIBUTION-FREE APPROACH FOR GROUPED SURVIVAL DATA: ANALYSIS AND CALCULATION OF EFFICIENCY

by

William H. Stewart

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Completed March 1979

Commencement June 1979

APPROVED:

Redacted for privacy

Professor of Statistics

in charge of major

Redacted for privacy

Chairman of Department of Statistics

Redacted for privacy

Dean of Graduate School

Date thesis is presented March 13, 1979

Typed by Charlene Fries for William H. Stewart

ACKNOWLEDGMENTS

My most sincere gratitude is extended to Dr. Donald A. Pierce, whose insightful direction and encouragement made this research possible. I also wish to thank Dr. David R. Thomas and Dr. David S. Birkes for their generosity and interest during the course of my graduate study. Special thanks is due my wife, Alice, who has shown continual patience and support throughout this endeavor.

This research was supported in part by United States Public Health Service grants ES-00040 and HL-16461.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

A DISTRIBUTION-FREE APPROACH FOR GROUPED SURVIVAL DATA: ANALYSIS AND CALCULATION OF EFFICIENCY

I. INTRODUCTION

A survival experiment is one in which the experimental information consists of observable response times such as time-to-death or time-to-failure. Survival experiments are usually carried out with a sample of n experimental units in such a way that n independent response times may be observed. In addition, some concomitant measurements, which the experimenter believes may be pertinent to survival time, are observed for each experimental unit. This additional information may be circumstantial depending on the various properties of the individual units, or it may consist of treatments which can be applied to units in a planned method, designed to study certain factors of interest. The primary statistical problem is to determine which factors of this concomitant information have important effects on survival time and to estimate these effects and the survival distributions that accrue from them.

One of the major fields of application for survival experiments is medical science. Gross and Clark (1975) give many such applications and examples. A common problem is to compare competing medical treatments for a disease. An experiment is designed so that concomitant information consists of an indication of which treatment a patient receives plus whatever additional factors are deemed important to survival. Then survival time, or time until some important response, is observed for each patient. The statistical problem is to detect differences in survival time caused by the competing treatments while adjusting for the other

factors. Also survival analysis is used in animal experiments, industrial life testing, and actuarial science.

A major complication for survival analysis is the problem of censored data. It is not always possible to observe the response time of each individual unit in an experiment. In most experiments this arises simply from the practical necessity of ending the experiment at some point in time at which some units may not yet have responded. Thus the information about a unit which survived the experiment is only that its response time was greater than the duration of the experiment. Also it may happen in certain types of experiments that the ability to monitor some units is lost during the experiment, so that in these cases the information available is only that the survival time is greater than the last point at which that unit was observed to be surviving. It is a major requirement, then, that a good method of statistical analysis for survival data be able to incorporate censored data in a simple, straightforward manner.

Of the more common parametric regression models employed in survival analysis are the exponential and Weibull models. Exponential regression models have been studied by Feigl and Zelen (1965), Zippin and Armitage (1966), Glasser (1967), and Prentice (1973). A larger class of models containing the exponential models are the Weibull regression models as given in the work of Peto and Lee (1973) and Prentice and Shillington (1975). These models have the advantages of providing computationally simple analysis and easily allowing for censored data; however, they restrict attention to a somewhat narrow class of distributions with monotone hazard functions. These models are often inadequate for explaining data derived from biological settings in which the hazard

functions depend on many factors and may increase or decrease on different time intervals. A possible approach to this problem is to use more general parametric models such as that proposed by Farewell and Prentice (1977), but these large parametric models are usually difficult to interpret and cumbersome to analyze.

Due to the difficulty in modeling survival time, nonparametric methods have been used extensively. The product limit estimator of Kaplan and Meier (1958) has often been used to estimate a survival curve. For comparing two survival distributions with censored data Gehan (1965) developed a generalized Wilcoxon test, and later Breslow (1970) extended this to a generalized Kruskal-Wallis test for censored k-sample problems. Mantel (1966) gave another rank statistic for comparing two survival distributions, while Peto and Peto (1972) and Peto (1972a) derive more general results by considering locally most powerful rank tests for Lehmann alternatives. These nonparametric methods are easy to apply and to interpret, but are limited to analysis of rather simple, single-factor experiments.

Cox (1972) developed a semi-nonparametric approach which generalized both the Weibull regression models and the nonparametric Lehmann alternatives. The form of an underlying hazard function was left unspecified with the covariables employed to have multiplicative effects on the underlying hazard. By reasoning conditionally on observed failure times Cox obtained a likelihood expression which did not depend on the underlying hazard. In a follow-up paper Cox (1975) termed this a "partial likelihood" and showed that for purposes of inference about regression parameters it could be treated as though it were a true likelihood function. The result of this was that one could use the partial likelihood

for inference about the effects of covariables on survival time without worrying about modeling the underlying hazard function. Furthermore, the statistical analysis was simple to apply, while the Cox models were rich enough to fit many different survival distributions and allowed for the use of flexible regression structures to analyze experiments with many covariables.

The motivation for the analysis to be developed here comes from consideration of data from experiments with laboratory animals. These experiments are conducted to determine the effects of toxicants and other stresses on the survival times of the animals. Focus will be directed to two special characteristics of these laboratory experiments throughout this study.

First, due to the controlled, laboratory conditions available in such experiments, it is possible to get replications of matched experimental units with exactly the same associated concomitant information. The treatments or associated information can ordinarily be chosen beforehand, so that the experiment may be designed to contain factorial combinations of the desired treatments. For example, if the joint effects of zinc and copper toxicants on survival rates of fish is under study, then factorial combinations of low, medium, and high concentrations of each toxicant might be of interest. Tanks could be prepared to give all the possible combinations of the two toxicants at these various concentration levels, and fish randomized to the tanks to give replications of each treatment combination.

The second distinctive trait of these laboratory experiments is that the exact failure time for each animal is generally not available. The failure times are grouped into convenient time periods, so that a

record is made only of the number of animals in each treatment combination which failed during each time period. In the example above the experiment might be run for ten days, with an observer counting the number of deaths in each tank at the end of each day.

Typically such toxicology data are not fit well by standard parametric regression models and often appear to involve mixtures of two or more distributions. Mixture models have been studied by Boag (1949), Berkson and Gage (1952), and Chen et al. (1977), but nonparametric methods seem more appropriate when the survival distributions are not well understood. The Cox model would appear to be flexible enough to fit most practical data sets of this sort; however, for grouped data Cox's conditional analysis does not provide a tractable partial likelihood. Peto (1972b) and Breslow (1972, 1974, 1975) suggest approximations for obtaining a likelihood function, but for heavily grouped data these methods are known to yield inconsistent estimators of the regression parameters.

For grouped data, such as will be considered here, Cox suggested a conditional logistic regression model which reduces to his continuous model as the grouping becomes finer and finer. Thompson (1977) has employed this model; however, there does not seem to be a convenient way to obtain a likelihood expression free of the underlying hazard, so that the analysis is burdened by computations involving nuisance parameters. Another approach for handling grouped data utilized by Prentice and Gloeckler (1978) reasons directly from the Cox continuous model, but again the analysis requires heavy computations with nuisance parameters.

The methods to be developed in this work are based on an approximation which allows for the formation of a maximum relative likelihood

function using Cox's model applied to grouped data. This likelihood is similar in form to the Cox partial likelihood and yields consistent estimators of the regression parameters. It presents an extremely flexible tool for data analysis, while allowing for a wide range of models to be fit and for standard likelihood inference procedures to be followed.

In addition, the approximation provides a convenient vehicle for examining the efficiency of the Cox semi-nonparametric method with other parametric procedures. The specification of nuisance parameters associated with the underlying hazard in the grouped data problems is a distinct advantage to efficiency studies made in the continuous data setting. The parametric models may be viewed as imposing certain smooth structures on the nuisance parameters, and the efficiency can be investigated directly by considering these restrictions. It is felt that the efficiency results for grouped data to be presented here provide an intuitive grasp of the nature of the Cox model.

II. ANALYSIS

II.1. The Cox Model

Consider a continuous failure time random variable T taking values on $(0,\infty)$. A realization of T may be thought of as the time-to-failure or time-to-response of an experimental unit. Let $F(t)$ denote the cumulative distribution function of T, and assume $\frac{d}{dt} F(t) = f(t)$ exists almost everywhere on $(0,\infty)$. It will be convenient to define $\bar{F}(t) = 1 - F(t)$ to be the survival function of T.

An important tool for survival analysis is the hazard function given by

$$\lambda(t) = \lim_{\Delta t \to o} \frac{Pr\{t \leq T < t + \Delta t \,|\, T \geq t\}}{\Delta t}$$

or $\lambda(t) = f(t)/\bar{F}(t)$, which exists a.s. Informally, one can think of $\lambda(t)$ as the instantaneous probability of failure at t, given survival up to t. The hazard function is defined in terms of the survival function, but we can also obtain the survival function from the hazard function, since we can write

$$\frac{-d \log \bar{F}(t)}{dt} = \lambda(t).$$

Then, integrating both sides,

$$\log \bar{F}(t) = -\int_o^t \lambda(u)du$$

or

$$\bar{F}(t) = \exp[-\int_o^t \lambda(u)du].$$

So either the survival function or the hazard function may be specified

to characterize the distribution of a continuous failure time random variable.

Suppose a collection $\{T_i : i = 1, \ldots, m\}$ of continuous failure time random variables, corresponding to m different treatments of a survival experiment, is considered. To explain treatment effects with a distribution free analysis, D. R. Cox (1972) proposed the regression model

$$\lambda_i(t) = \lambda_o(t) \exp\{\beta' x_i(t)\}, \tag{1.1}$$

where $\lambda_i(t)$ is the hazard function for T_i, $\lambda_o(t)$ is an arbitrary, unknown hazard function, β is a p-dimensional vector of parameters, and $x_i(t)$ is a p-dimensional vector of covariables, possibly dependent on time, which describes treatment i. This models the treatments to have multiplicative effects on an underlying hazard $\lambda_o(t)$ which is allowed to be any continuous time hazard function. The model is nonparametric or distribution-free in the sense that the underlying hazard is unrestricted; however, the treatment effect is expressed by a parametric regression expression.

If x_i is not allowed to be time dependent, then the Cox model is the proportional hazards model where the treatment effect is to multiply the underlying hazard by the constant $\exp(\beta' x_i)$. However, allowing the covariables to be time dependent produces much richer classes of models with the possibility that the multiplicative treatment effects may increase or decrease with time. In many applications, especially those related to biological phenomena, the treatment effects do appear to depend quite heavily on time. By careful choice of time dependent

variables, it is often possible to obtain a good fit for a survival data set with the Cox model.

Besides the intuitive appeal of the multiplicative treatment effects in this model, Cox has given a very neat analysis for continuous-time data which avoids parameterizing $\lambda_o(t)$ and allows standard likelihood procedures to be carried out for estimation and hypothesis tests of β. For the grouped data problems to be considered in this work the essential Cox model is retained, and an analysis is developed which leaves $\lambda_o(t)$ free while allowing for direct inference about the regression parameters.

II.2. The Grouped Data Model

The survival experiments of concern here consist of m different treatments with n_i, $i = 1, \ldots, m$, individuals assigned to each treatment. The treatments are to be thought of in a broad sense, encompassing any information that the experimenter can use to distinguish experimental units, although typically the treatment structure will be a factorial design for studying certain effects of interest. For each individual a measurement is made of either response time or censoring time, that is, the time survived until removal from the experiment. It is assumed that the experiment is conducted in such a way so that the survival times of all individuals are independent. Also survival times for individuals with the same treatment will be assumed identically distributed. Additionally, no relation between response time and censoring time is supposed. Thus for censored observations only the information that the response time must be greater than the censoring time is used.

The focus here is on experiments in which survival time is grouped into a finite number of time periods. Let the time axis be partitioned by positive real numbers $a_1, \ldots, a_k$, and define $a_o = 0$. The experiment is conducted for k time periods with time period j, $j = 1, \ldots, k$, represented by the time interval $(a_{j-1}, a_j]$. At the end of each time period observations are taken on each experimental unit to determine if that unit failed, survived, or was censored during that time period.

Again let T_i denote the response-time random variable for an individual with treatment i. There are three types of events which may be observed for an individual with treatment i: (1) response during some j^{th} period, (2) censorship during some j^{th} period, or (3) survival of the experiment. These events may be described respectively as (1) $a_{j-1} < T_i \leq a_j$, (2) $T_i > a_{j-1}$, and (3) $T_i > a_k$. Thus if the probability distribution of T_i is known, then the probabilities of the possible observed events may be calculated directly.

It will be convenient to define parameters which essentially describe the discrete time hazard functions for the m treatments. Let

$$p_{ij} = \Pr\{a_{j-1} < T_i \leq a_j \mid T_i > a_{j-1}\},$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, k$. Then p_{ij} is the conditional probability of failing in period j given survival of j-1 periods under treatment i. Likewise define

$$q_{ij} = 1 - p_{ij} = \Pr\{T_i > a_j \mid T_i > a_{j-1}\},$$

so q_{ij} is the conditional probability of surviving the j^{th} period given survival of j-1 periods under treatment i. The probabilities for the possible observed events for the individuals may be expressed as follows.

(1) Response in period j:

$$\Pr\{a_{j-1} < T_i \leq a_j\} = \left(\prod_{\ell=1}^{j-1} q_{i\ell}\right) p_{ij}.$$

(2) Censorship in period j:

$$\Pr\{T_i > a_{j-1}\} = \prod_{i=1}^{j-1} q_{i\ell}.$$

(3) Survival of the experiment:

$$\Pr\{T_i > a_k\} = \prod_{j=1}^{k} q_{ij}.$$

Since the survival times of all individuals are independent, the likelihood function is obtained by the product of the probabilities of all observed individual events. To write the likelihood simply, let r_{ij} be the number in treatment i which respond during period j, and let s_{ij} be the number in treatment i which survive period j. If an individual is censored during period j, it will not be considered to have responded in or survived period j. Note from expressions (1), (2), and (3) above that for an individual with treatment i, if it survives period j it contributes the factor q_{ij} to the likelihood, while if it responds during period j it contributes the factor p_{ij}. Thus the likelihood may be obtained by counting the numbers responding and surviving at each time period. We then have

$$L(p) \propto \prod_{i=1}^{m} \prod_{j=1}^{k} p_{ij}^{r_{ij}} q_{ij}^{s_{ij}} \tag{2.1}$$

where p is the km-dimensional vector of the p_{ij}.

A successful analysis of this grouped data problem involves a modeling of the p_{ij} to take advantage of information about the treatments. The model employed here follows a suggestion of Kalbfleisch and Prentice (1973). The continuous Cox model (1.1) is assumed, and then from this a modeling for the conditional grouped data parameters q_{ij} is obtained. Additionally it is assumed that the covariable $x_i(t)$ is constant on each interval $(a_{j-1}, a_j]$. The value of $x_i(t)$ on $(a_{j-1}, a_j]$ will be written x_{ij}. Then applying the Cox model

$$
\begin{aligned}
q_{ij} &= \Pr\{T_i > a_j \mid T_i > a_{j-1}\} \\
&= \frac{\bar{F}_i(a_j)}{\bar{F}_i(a_{j-1})} \\
&= \frac{\exp[-\int_o^{a_j} \lambda_i(t)dt]}{\exp[-\int_o^{a_{j-1}} \lambda_i(t)dt]} \\
&= \exp\{-\int_{a_{j-1}}^{a_j} \lambda_o(t)\, \exp[\beta' x_i(t)]dt\} \\
&= \exp\{-\lambda_j \exp(\beta' x_{ij})\}, \qquad (2.2)
\end{aligned}
$$

where $\lambda_j = \int_{a_{j-1}}^{a_j} \lambda_o(t)dt$.

This produces a model for the q_{ij} with k parameters $\lambda = (\lambda_1, \ldots, \lambda_k)'$ associated with the underlying hazard function and p parameters $\beta = (\beta_1, \ldots, \beta_p)'$ associated with the treatment effects. The problem is then reduced from one with km unknown parameters q_{ij} to one with the k + p parameters of the model. As in the continuous model the underlying hazard is still unrestricted, since the λ_j are free to take on any positive values that the integrals of $\lambda_o(t)$ might be. The model is then

partially nonparametric, since the underlying hazard is left completely free, but the choice of the x_{ij} will restrict the model to a certain set of parametric alternatives from the arbitrary $\lambda_o(t)$. However, by liberal use of the time dependent covariables, classes of alternatives can be made large enough to encompass most practical problems.

Using (2.2) and the equation developed for the likelihood in (2.1), the likelihood function for the model can be written as

$$L(\beta,\lambda) \propto \prod_{i=1}^{m} \prod_{j=1}^{k} \{1-\exp[-\lambda_j \exp(\beta' x_{ij})]\}^{r_{ij}} \{\exp[-\lambda_j \exp(\beta' x_{ij})]\}^{s_{ij}}.$$

Writing $\ell(\beta,\lambda) = \log L(\beta,\lambda)$ and dropping constants,

$$\ell(\beta,\lambda) = \sum_{i=1}^{m} \sum_{j=1}^{k} \{r_{ij} \log[1-\exp\{-\lambda_j \exp(\beta' x_{ij})\}] - s_{ij}[\lambda_j \exp(\beta' x_{ij})]\}. \tag{2.3}$$

This is essentially the log likelihood function used by Prentice and Gloeckler (1978). Their approach was to use maximum likelihood methods on (2.3), proceeding to estimate both β and λ by iterative methods. For problems in which k is large and the interest is on β, working with (2.3) directly involves difficult computations to estimate the nuisance parameters λ in conjunction with the parameters of interest β. To circumvent this problem the following sections show how certain good approximations to (2.3) may be made which allow the construction of a likelihood expression which depends only on β. Then, much in the spirit of the Cox partial likelihood, analysis about β can proceed free of the nuisance parameters λ.

II.3. The Approximate Likelihood

To develop a workable approximation to (2.3) the following assumptions are added to the grouped data problem presented in the previous section. (i) Assume $\lambda_i(t)$ to be constant on each time period, in particular let $\lambda_o(t) = h_j$ and $\lambda_i(t) = h_j \exp(\beta' x_{ij})$, for $a_{j-1} < t \leq a_j$. (ii) Assume that for observed failures the exact failure times are known. If R_{ij} is the set of all individuals with treatment i which failed during time period j, then let $t_{ij\ell}$ be the exact failure time for an individual $\ell \in R_{ij}$. (iii) Assume exact censoring times are at the a_j.

Assumption (i) is very weak; in fact, the model for the grouped data parameters q_{ij} adding (i) only does not differ from (2.2). Assumption (ii), at first glance, appears to violate the whole structure of the grouped data problem. However, what is intended is to select pseudo-failure times $t_{ij\ell}$ in such a way as to be consistent with the observed grouped data. The pseudo-failure times will then provide a vehicle for obtaining a good approximate likelihood for the actual grouped data problem.

Again let the integrated hazard parameters be defined as

$$\lambda_j = \int_{a_{j-1}}^{a_j} \lambda_o(t)dt = (a_j - a_{j-1})h_j.$$

Also write $c_{ij\ell} = (t_{ij\ell} - a_{j-1})/(a_j - a_{j-1})$ for the proportion of the j^{th} interval survived by individual ℓ of R_{ij}. Using (i) and (ii) the following expressions are obtained for the conditional probabilities of survival and failure.

$$\Pr(T_i > a_j \mid T_i > a_{j-1}) = \exp[-\lambda_j \exp(\beta' x_{ij})],$$

$$f_{T_i|T_i>a_{j-1}}(t_{ij} \mid T_i > a_{j-1}) \propto \lambda_j \exp(\beta' x_{ij}) \exp[-c_{ij\ell}\lambda_j \exp(\beta' x_{ij})]$$

The likelihood function is then formed as the product of the conditional probabilities of the observed survivals and failures.

$$L_c(\lambda,\beta) = \prod_{i=1}^{m} \prod_{j=1}^{k} \left(\prod_{\ell \in R_{ij}} \{\lambda_j \exp(\beta' x_{ij}) \exp[-c_{ij\ell} \lambda_j \exp(\beta' x_{ij})]\} \right.$$

$$\left. \cdot \{\exp[-\lambda_j \exp(\beta' x_{ij})]\}^{s_{ij}} \right].$$

Writing $c_{ij} = \left(\sum_{\ell \in R_{ij}} c_{ij\ell} \right)/r_{ij}$, this becomes

$$L_c(\lambda,\beta) = \prod_{i=1}^{m} \prod_{j=1}^{k} [\{\lambda_j \exp(\beta' x_{ij})\}^{r_{ij}} \exp\{-c_{ij} r_{ij} \lambda_j \exp(\beta' x_{ij})\}$$

$$\cdot \{\exp[-\lambda_j \exp(\beta' x_{ij})]\}^{s_{ij}}].$$

The log likelihood is then

$$\ell_c(\lambda,\beta) = \sum_{i=1}^{m} \sum_{j=1}^{k} \{r_{ij} \log \lambda_j + r_{ij} \beta' x_{ij}$$

$$- (s_{ij} + c_{ij} r_{ij}) \lambda_j \exp(\beta' x_{ij})\}. \tag{3.1}$$

This likelihood expression is similar in form to those used by Breslow (1974) and Holford (1977). Note that to form this log likelihood it is not necessary to select pseudo-failure times for each individual, but only to select a pseudo-value for each c_{ij}. Since c_{ij} is the average proportion of the j^{th} interval survived by those failing in interval j under treatment i, it is natural to consider the conditional expectation

$$E\{\frac{T_i - a_{j-1}}{a_j - a_{j-1}} \mid a_{j-1} < T_i \leq a_j\}.$$

Using assumption (i) with the Cox model, the conditional density of $Y_{ij} = (T_i - a_{j-1})/(a_j - a_{j-1})$ given $a_{j-1} < T_i \leq a_j$, i.e., $0 < Y_{ij} \leq 1$, is found to be the truncated exponential density

$$f_{Y_{ij}}(y) = \frac{\phi_{ij}\exp(-\phi_{ij}y)}{1 - \exp(-\phi_{ij})},$$

$0 < y \leq 1$, with $\phi_{ij} = \lambda_j \exp(\beta' x_{ij})$.

It follows then that

$$E(Y_{ij}) = \int_0^1 y \frac{\phi_{ij}\exp(-\phi_{ij}y)}{1 - \exp(-\phi_{ij})}\, dy$$

$$= \frac{1}{\phi_{ij}} - \frac{\exp(-\phi_{ij})}{1 - \exp(-\phi_{ij})}$$

Recall from (2.2) that $q_{ij} = \exp(-\phi_{ij})$, so

$$(3.2) \qquad E\{\frac{T_i - a_{j-1}}{a_j - a_{j-1}} \mid a_{j-1} < T_i \leq a_j\} = -\frac{1}{\log q_{ij}} - \frac{q_{ij}}{1 - q_{ij}}.$$

Using expression (3.2) for the values of c_{ij} would be an ideal choice; however, since the q_{ij} are unknown parameters, this is not possible. When the q_{ij} are large, Holford (1976) suggests using $c_{ij} = \frac{1}{2}$ for all i and j. This can be gotten by noting the limiting value for (3.2) as $q_{ij} \to 1$ is $\frac{1}{2}$, or recognizing that the failure time given failure in the j^{th} interval is well approximated by a uniform distribution when q_{ij} is large. In general, though, expression (3.2) is less than $\frac{1}{2}$, so that the use of $c_{ij} = \frac{1}{2}$ everywhere will slightly bias estimators based on (3.1). For this approximation to work well the experimenter would be

required to choose intervals which keep all q_{ij} large, but this cannot always be foreseen or controlled.

Advantage can be made of the n_i replications available for each treatment in such experiments. A naive estimator of q_{ij} without considering modeling can be formed by

$$\hat{q}_{ij} = \frac{s_{ij}}{r_{ij} + s_{ij}},$$

the proportion of survivors of the total number at risk throughout the interval. Then the c_{ij} may be estimated by

$$\hat{c}_{ij} = -\frac{1}{\log \hat{q}_{ij}} - \frac{\hat{q}_{ij}}{1 - \hat{q}_{ij}}. \tag{3.3}$$

If $r_{ij} = 0$, then $\hat{c}_{ij} = .5$ is used, and if $s_{ij} = 0$, then $\hat{c}_{ij} = 0$ is used, as these are the limiting values of c_{ij} for $q_{ij} \to 1$ and $q_{ij} \to 0$, respectively.

Asymptotic considerations make $\hat{c}_{ij}$ an especially appealing choice for c_{ij}. As long as the censoring mechanism is not too severe and large sample size embodies each n_i being large, then the $\hat{c}_{ij}$ will converge to c_{ij} as sample size gets large. It is proposed, then, that the true log likelihood expression of (2.3) be replaced by the log likelihood of (3.1) with the c_{ij} approximated by the $\hat{c}_{ij}$ of (3.3). The approximate log likelihood may then be written as

$$\ell_{\hat{c}}(\beta,\lambda) = \sum_{i=1}^{m} \sum_{j=1}^{k} \{r_{ij} \log \lambda_j + r_{ij}\beta' x_{ij} - (s_{ij} + \hat{c}_{ij} r_{ij})\lambda_j \exp(\beta' x_{ij})\}. \tag{3.4}$$

The key to analysis with (3.4) is the formation of a maximum relative likelihood function $\ell^*(\beta) = \max_{\lambda} \ell_{\hat{c}}(\beta,\lambda)$ which effectively removes the nuisance parameters associated with the underlying hazard. For fixed β let $\hat{\lambda}(\beta)$ be the value of λ that maximizes $\ell_{\hat{c}}(\beta,\lambda)$. The advantage of using $\ell_{\hat{c}}$ instead of the true ℓ is that $\hat{\lambda}(\beta)$ can be solved explicitly in terms of β and the data. Solving for the maximizing value gives

$$\hat{\lambda}_j(\beta) = r_{\cdot j}/\{\sum_{i=1}^{m} (s_{ij} + \hat{c}_{ij} r_{ij}) \exp(\beta' x_{ij})\}, \tag{3.5}$$

where $r_{\cdot j} = \sum_{i=1}^{m} r_{ij}$. $\ell^*(\beta)$ can then be written as

$$\begin{aligned} \ell^*(\beta) &= \ell_{\hat{c}}(\beta, \hat{\lambda}(\beta)) \\ &= \sum_{j=1}^{k} \{ \sum_{i=1}^{m} r_{ij} \beta' x_{ij} - r_{\cdot j} \log[\sum_{i=1}^{m} (s_{ij} + \hat{c}_{ij} r_{ij}) \exp(\beta' x_{ij})] \} + C, \end{aligned} \tag{3.6}$$

with $C = \sum_{j=1}^{k} (r_{\cdot j} \log r_{\cdot j} - r_{\cdot j})$. For purposes of inference about the regression parameters β, $\ell^*(\beta)$ may be used exactly as if it were a log likelihood function for β.

The maximum likelihood estimates $\hat{\beta}$ and $\hat{\lambda}$, i.e., the values that maximize $\ell_{\hat{c}}(\beta,\lambda)$, may be obtained simply by maximizing $\ell^*(\beta)$. If $\hat{\beta}$ maximizes $\ell^*(\beta)$, then $\hat{\beta}$ and $\hat{\lambda}(\hat{\beta})$ will maximize $\ell_{\hat{c}}(\beta,\lambda)$. This follows by noting

$$\max_{\beta} \ell^*(\beta) = \max_{\beta} \{ \max_{\lambda} \ell_{\hat{c}}(\beta,\lambda) \} = \max_{\beta,\lambda} \ell_{\hat{c}}(\beta,\lambda).$$

Furthermore, if β is restricted to some subspace $\Omega_o \subset R^p$ and λ left free, the restricted maximum likelihood estimates $\hat{\beta}_o$ and $\hat{\lambda}_o$ can still be found with $\ell^*(\beta)$. Again if $\hat{\beta}_o$ is the value that maximizes $\ell^*(\beta)$ for $\beta \in \Omega_o$,

then $\hat{\beta}_o$ and $\hat{\lambda}(\hat{\beta}_o)$ are the values that maximize $\ell_{\hat{c}}(\beta,\lambda)$ for $\beta \in \Omega_o$. Maximizing $\ell^*(\beta)$, with or without restrictions, by the Newton-Raphson algorithm presents no difficulties, since the dimension of β is usually small. Typically great savings in calculations are made by using $\ell^*(\beta)$ instead of $\ell(\beta,\lambda)$, as the dimension of λ is likely to be large for a well designed experiment.

Large sample, asymptotic analysis may be carried out following the lines of Cox and Hinkley (1974, Chapter 9). To apply this theory the asymptotic normality of the maximum likelihood estimators needs to be established. Assuming exact failure times, $\ell_{\hat{c}}(\beta,\lambda)$ is the log likelihood function and is formed with independent observations coming from the m different populations corresponding to the m treatments. Bradley and Gart (1962) have extended the asymptotic theory of maximum likelihood estimators to such cases. These authors require the following conditions to establish asymptotic normality:

(i) existence of partial derivatives of $\ell(\beta,\lambda)$ up through 3^{rd} order;

(ii) interchange of the order of differentiation and integration;

(iii) positive definite matrix of expectations of second partial derivatives of $\ell(\beta,\lambda)$;

(iv) $N \to \infty$, so that $n_i = \mu_i N$, where $0 < \mu_i < 1$ for $i = 1, \ldots, m$, and $\sum_{i=1}^{m} \mu_i = 1$.

For reasonable censoring mechanisms these conditions apply here, and $\hat{\theta} = \begin{bmatrix} \hat{\beta} \\ \hat{\lambda} \end{bmatrix}$ has approximately a multivariate normal distribution with mean equal to the true value of $\theta = \begin{bmatrix} \beta \\ \lambda \end{bmatrix}$ and variance matrix approximated by $i^{-1}(\hat{\beta},\hat{\lambda})$, where $i(\hat{\beta},\hat{\lambda})$ is the observed Fisher information, i.e., the matrix of second partials of $\ell_{\hat{c}}(\beta,\lambda)$ evaluated at $\hat{\beta}$ and $\hat{\lambda}$.

$\ell^*(\beta)$ can be conveniently employed in the likelihood ratio test. Let $\beta = (\beta_1, \beta_2, \ldots, \beta_p)' \in R^p$, and suppose we consider an hypothesis test of the form $H_o: \beta \in \Omega_o$ vs $H_a: \beta \notin \Omega_o$, where $\Omega_o = \{\beta: \beta_1 = \beta_1^o, \ldots, \beta_r = \beta_r^o\}$, $r \leq p$. Then the test can be based on the statistic

$$
\begin{aligned}
W &= 2\{\sup_{\beta,\lambda} \ell_{\hat{c}}(\beta,\lambda) - \sup_{\beta\in\Omega_o,\lambda} \ell_{\hat{c}}(\beta,\lambda)\} \\
&= 2\{\sup_{\beta} \ell^*(\beta) - \sup_{\beta\in\Omega_o} \ell^*(\beta)\} \\
&= 2\{\ell^*(\hat{\beta}) - \ell^*(\hat{\beta}_o)\},
\end{aligned}
$$

where $\hat{\beta}$ and $\hat{\beta}_o$ are the unrestricted and restricted maximum likelihood estimators, respectively. Using the asymptotic normality of the maximum likelihood estimators and applying an argument such as given in Cox and Hinkley (1974, pp. 322-323), the limiting distribution of W is chi-square with r degrees of freedom. The distribution is central chi-square when H_o is true.

In comparing various regression models, one can form an analysis of deviation table analogous to an analysis of variance table, except twice the deviation in ℓ^* is used in place of reduction sums of squares. Also a goodness-of-fit test of a given regression model can be constructed using the approximate likelihood expressions. Suppose a "saturated" model is fit with the q_{ij} completely unrestricted, or equivalently, the ϕ_{ij} unrestricted. Then (3.4) can be rewritten as

$$
\ell_{\hat{c}}(\phi) = \sum_{i=1}^{m} \sum_{j=1}^{k} \{r_{ij} \log \phi_{ij} - (s_{ij} + \hat{c}_{ij} r_{ij})\phi_{ij}\}.
$$

This expression is maximized by $\hat{\phi}$, where $\hat{\phi}_{ij} = r_{ij}/(s_{ij} + \hat{c}_{ij} r_{ij})$. So

$2\{\ell_{\hat{C}}(\hat{\phi}) - \ell^*(\hat{\beta})\}$ has an approximate central chi-square distribution with $mk - p - k$ degrees of freedom when the regression model is adequate.

Once an appropriate regression model has been chosen, confidence regions for the parameters may be found neatly by taking advantage of $\ell^*(\beta)$. Let

$$i(\beta,\lambda) = \begin{bmatrix} i_{\beta\beta}(\beta,\lambda) & i_{\beta\lambda}(\beta,\lambda) \\ i_{\lambda\beta}(\beta,\lambda) & i_{\lambda\lambda}(\beta,\lambda) \end{bmatrix}$$

$$= -\begin{bmatrix} \frac{\partial^2 \ell_{\hat{C}}(\beta,\lambda)}{\partial\beta^2} & \frac{\partial^2 \ell_{\hat{C}}(\beta,\lambda)}{\partial\beta\partial\lambda} \\ \frac{\partial^2 \ell_{\hat{C}}(\beta,\lambda)}{\partial\lambda\partial\beta} & \frac{\partial^2 \ell_{\hat{C}}(\beta,\lambda)}{\partial\lambda^2} \end{bmatrix}$$

An estimate of the variance matrix for $\hat{\beta}$ and $\hat{\lambda}$ is given by $i^{-1}(\hat{\beta},\hat{\lambda})$. Adopt the following notation:

$$i^*(\beta) = \frac{-\partial^2 \ell^*(\beta)}{\partial\beta^2}, \quad \text{and} \quad \delta(\beta) = \frac{\partial\hat{\lambda}(\beta)}{\partial\beta}.$$

Richards (1961) has established the following facts:

$$\delta(\hat{\beta}) = -i_{\lambda\lambda}^{-1}(\hat{\beta},\hat{\lambda})\ i_{\lambda\beta}(\hat{\beta},\hat{\lambda}),$$

and

$$i^*(\hat{\beta}) = i_{\beta\beta}(\hat{\beta},\hat{\lambda}) - i_{\beta\lambda}(\hat{\beta},\hat{\lambda})\ i_{\lambda\lambda}^{-1}(\hat{\beta},\hat{\lambda})\ i_{\lambda\beta}(\hat{\beta},\hat{\lambda}).$$

Also note that for a partitioned, invertible, symmetric matrix of the form $\begin{bmatrix} A & C \\ C' & B \end{bmatrix}$, with $D = A - CB^{-1}C'$, that

$$\begin{bmatrix} A & C \\ C' & B \end{bmatrix}^{-1} = \begin{bmatrix} D^{-1} & -D^{-1}CB' \\ -B^{-1}C'D^{-1} & B^{-1} + B^{-1}C'D^{-1}CB^{-1} \end{bmatrix}$$

This fact, coupled with Richards' results, enables us to write

$$i^{-1}(\hat{\beta},\hat{\lambda}) = \begin{bmatrix} \{i^*(\hat{\beta})\}^{-1} & \{i^*(\hat{\beta})\}^{-1}\delta'(\hat{\beta}) \\ \delta(\hat{\beta})\{i^*(\hat{\beta})\}^{-1} & i_{\lambda\lambda}^{-1}(\hat{\beta},\hat{\lambda}) + \delta(\hat{\beta})\{i^*(\hat{\beta})\}^{-1}\delta'(\hat{\beta}) \end{bmatrix}.$$

The variance matrix for $\hat{\beta}$ and $\hat{\lambda}$ is then expressible in terms of $\{i^*(\hat{\beta})\}^{-1}$, $\delta(\hat{\beta})$, and $i_{\lambda\lambda}^{-1}(\hat{\beta},\hat{\lambda})$. All of these matrices are easy to compute. The first two come directly from the simple expressions for $\ell^*(\beta)$ and $\hat{\lambda}(\beta)$, whereas $i_{\lambda\lambda}^{-1}(\hat{\beta},\hat{\lambda})$ can be readily computed, since $i_{\lambda\lambda}(\hat{\beta},\hat{\lambda})$ is a diagonal matrix. Confidence intervals for β and λ may then be constructed using the asymptotic normality of $\hat{\beta}$ and $\hat{\lambda}$.

Finally, to estimate the survival curve for a given treatment combination one can use the relation

$$\begin{aligned} \bar{F}_i(a_j;\theta) &= \prod_{\ell=1}^{j} q_{i\ell} \\ &= \prod_{\ell=1}^{j} \exp\{-\lambda_\ell \exp(\beta' x_{i\ell})\} \\ &= \exp\{-\sum_{\ell=1}^{j} \lambda_\ell \exp(\beta' x_{i\ell})\}. \end{aligned}$$

Then the maximum likelihood estimator for $\bar{F}_i(a_j;\theta)$ is

$$\bar{F}(a_j;\hat{\theta}) = \exp\{-\sum_{\ell=1}^{j} \hat{\lambda}_\ell \exp(\hat{\beta}' x_{i\ell})\}.$$

To obtain confidence intervals for $\bar{F}(a_j;\theta)$ it is perhaps best to follow a suggestion of Prentice and Gloeckler (1978) and work with

$Y(\hat{\theta}) = \log\{-\log \bar{F}(a_j;\hat{\theta})\}$. The transformation to Y is preferable to $\bar{F}(a_j;\hat{\theta})$ for normal approximations, since it removes the range restrictions. The distribution of $Y(\hat{\theta})$ may be approximated by a normal distribution with mean $Y(\theta)$ and variance given by $\sigma^2 = [g(\hat{\theta})]'[i^{-1}(\hat{\theta})][g(\hat{\theta})]$, where $g(\theta) = \partial Y(\theta)/\partial\theta$. Then an approximate $(1-\alpha)$ level confidence interval for $Y(\theta)$ is given by

$$(Y(\hat{\theta}) - z_\alpha\sigma,\ Y(\hat{\theta}) + z_\alpha\sigma),$$

where z_α is the value such that a standard normal variate falls within $-z_\alpha$ and z_α with probability $1-\alpha$. Transforming the confidence interval to one for $\bar{F}(a_j;\theta)$ gives

$$([\bar{F}(a_j;\hat{\theta})]^{\exp(z_\alpha\sigma)},\ [\bar{F}(a_j;\hat{\theta})]^{\exp(-z_\alpha\sigma)})$$

as a $1-\alpha$ level confidence interval for $\bar{F}(a_j;\theta)$. Confidence bands for the survival curve may be constructed by considering the confidence intervals of $\bar{F}(a_j;\theta)$ for each j.

II.4. A Toxicology Example

An experiment was carried out by Garton (1975) to study the effects of either one week or two weeks acclimation time on the survival rate of fish subjected to zinc toxicants. Three levels of zinc concentration were used, and a 2×3 factorial experiment was run on the two acclimation periods with the zinc concentrations. There were 50 fish randomized to each of the six treatment combinations with mortality observed on a daily basis for ten days. Time here was measured from the introduction of the zinc toxicants after the desired acclimation period had been established. Table 1 gives the observed daily mortalities.

Table 1. Daily mortality (from groups of fifty fish per treatment combination)

Acclimation Time:	One Week			Two Weeks		
Zinc Concentration:	Low	Med.	High	Low	Med.	High
Day/Mortality						
1	0	0	0	0	0	0
2	3	3	2	0	1	3
3	12	17	22	13	21	24
4	11	16	15	8	8	10
5	3	5	7	0	5	4
6	0	1	1	0	0	1
7	0	0	2	0	0	0
8	0	1	0	0	0	0
9	0	0	0	0	0	0
10	0	0	0	0	0	0

Days 1, 2 and 8, 9, 10 were each combined giving k = 7 time periods. It was decided initially to consider a regression on six covariables. Letting A represent acclimation, C represent linear log concentration of zinc, and T represent linear time, C, C × T, A, A × T, A × C, and C × T^2 were selected as covariables. The variable A was coded 0 for one week acclimation and 1 for two weeks acclimation. The three zinc concentrations were in ratio 67:100:128. The variable C was coded by 0.2047, 0.6052, 0.8520, which were 4 less than the logarithms of the above ratio figures. The variable T was coded as -3, -2, -1, 0, 1, 2, 3. Thus for the six treatment combinations and seven time periods the model was

$$q_{ij} = \exp\{-\lambda_j \exp(\beta' x_{ij})\},$$

$i = 1, \ldots, 6$, $j = 1, \ldots, 7$, where $[x_{ij}]' = (C, C \times T, A, A \times T, A \times C, C \times T^2)$. For example, for the treatment combination of two weeks acclimation with medium zinc concentration, say $i = 5$, and for the sixth time period (the seventh day) the model gave

$$q_{56} = \exp\{-\lambda_6 \exp[(.6052)\beta_1 + (.6052 \times 2)\beta_2 + (1)\beta_3 + (1 \times 2)\beta_4 + (1 \times .6052)\beta_5 + (.6052 \times 4)\beta_6]\}.$$

Using analysis based on $\ell^*(\beta)$, the following analysis of deviation table was obtained.

Table 2. Analysis of deviation

Source	d.f.	$\chi^2 = 2\Delta\ell^*$
C	1	37.42
C × T	1	9.32
A effects	3	15.69
A	1	4.68
A × T	1	10.27
A × C	1	.74
$C \times T^2$	1	2.97
Lack of fit	29	22.51

The χ^2 statistics were obtained by successively adding parameters to the model. For example, the χ^2 statistic 9.32 opposite C × T in the table was the test for $\beta_2 = 0$ with β_1 in the model, i.e.,

$$9.32 = 2[\ell^*(\hat{\beta}_1, \hat{\beta}_2, 0, 0, 0, 0) - \ell^*(\hat{\hat{\beta}}, 0, 0, 0, 0, 0)],$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ were the restricted MLE's for $\beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ and $\hat{\hat{\beta}}_1$ was the restricted MLE for $\beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$. The lack of fit

statistic was obtained by fitting a saturated model as given in section II.3. The statistic was $\chi^2 = 2[\ell_{\hat{c}}(\hat{\phi}) - \ell^*(\hat{\beta})]$ with $mk - p - k = 29$ degrees of freedom.

Examination of Table 2 reveals that the data can be fit very nicely by just the four variables C, C × T, A, and A × T. The lack of fit statistic corresponding to this model is $\chi^2 = .74 + 2.97 + 22.51 = 26.22$ on 31 degrees of freedom. $\ell^*(\beta)$ was maximized with respect to this model to obtain estimates of β, and $[i^*(\hat{\beta})]^{-1}$ used for estimates of standard errors. Once β was estimated, estimates of λ were obtained by using expression (3.5) for $\hat{\lambda}_j(\hat{\beta})$. Tables 3 and 4 give the estimates of β and λ.

Table 3. Estimates of β

Term	$\hat{\beta}$	Standard Error
C	3.01	.53
C × T	.98	.30
A	-.98	.26
A × T	-.48	.15

Table 4. Estimates of λ

$\hat{\lambda}_1 = .030$

$\hat{\lambda}_2 = .264$

$\hat{\lambda}_3 = .192$

$\hat{\lambda}_4 = .076$

$\hat{\lambda}_5 = .008$

$\hat{\lambda}_6 = .004$

$\hat{\lambda}_7 = .001$

Table 4 illustrates the humped shape of the underlying hazard function for this problem. Analysis with constant hazard or monotone hazard models would not be appropriate here. Survival functions for the treatments were then estimated using the relation

$$\bar{F}(a_j;\hat{\theta}) = \exp\{-\sum_{\ell=1}^{j} \hat{\lambda}_\ell \exp(\hat{\beta}'x_{i\ell})\}.$$

Figure 1 gives a graph of all six survival curves as estimated by the above relation.

The effect of the zinc concentration was the strongest factor in explaining the data. This effect does get stronger with time, so that the C and C × T variables were both needed to fit the data well. There was also a marked effect due to acclimation time, even though during the first three days there was practically no difference in survival rates between one-week and two-week acclimations. The one-week groups survived at slightly higher rates in the early time periods, but this difference was attributable to sampling error. After three days the fish acclimated for two weeks did remarkably better than those with one week of acclimation time, the difference becoming greater with time. It was imperative to have the time dependent variable A × T to explain this effect of acclimation adequately. It appears that the benefits of an extended acclimation period do not come into play until the fish have been subjected to the toxicant for a lengthy period of at least three days.

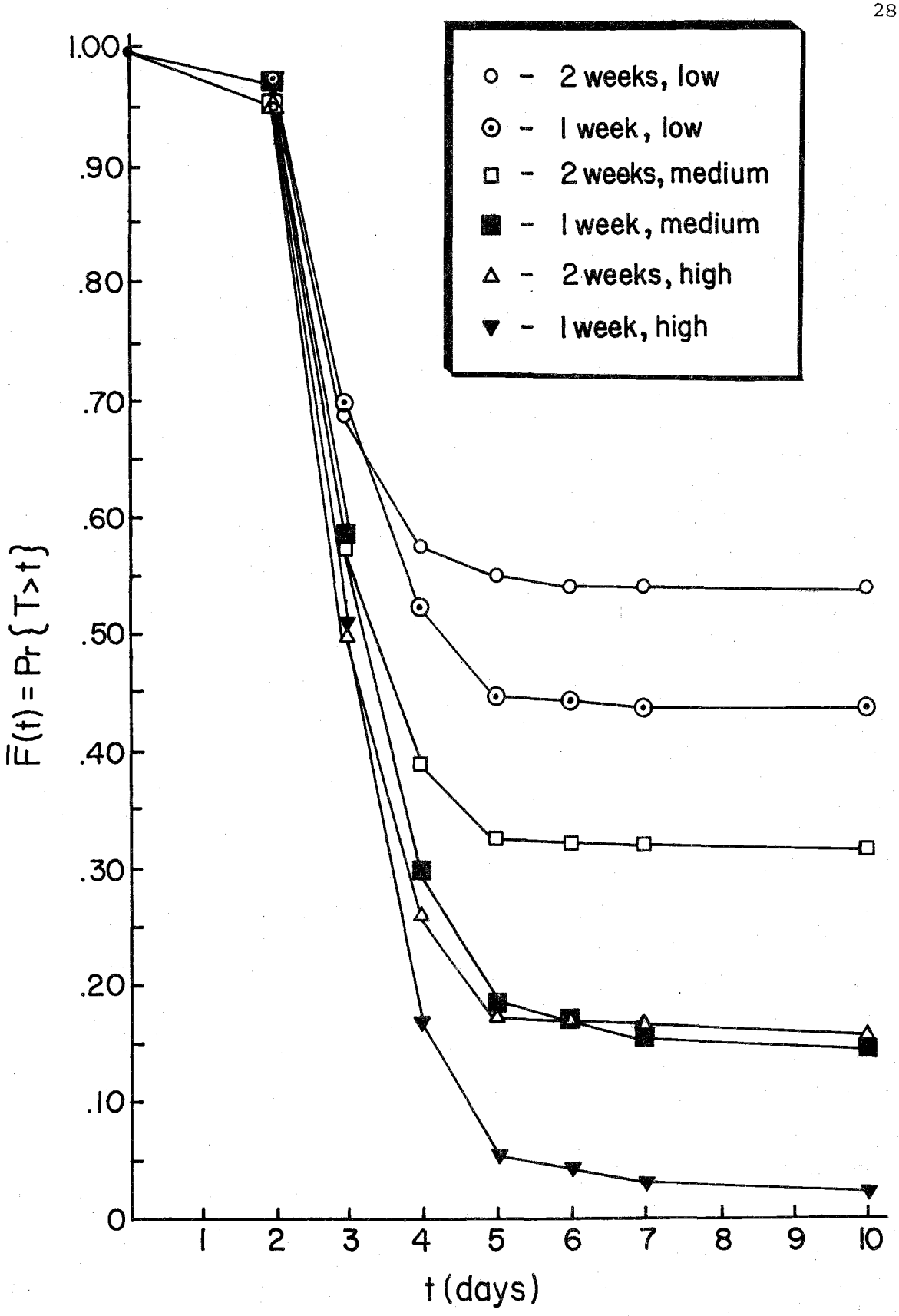

Figure 1. Estimated survival curves

II.5. An Alternative Approach to the Approximation

Recall the following notation. Let $\theta = \begin{bmatrix}\beta\\ \lambda\end{bmatrix}$, and let $\ell(\theta)$ denote the true likelihood of (2.3), while $\ell_c(\theta)$ denotes the approximate likelihood given in (3.1) with some fixed choice of c, the vector of all c_{ij}. Letting $\phi_{ij} = \lambda_j \exp(\beta' x_{ij})$, it is seen that the approximate likelihood may be obtained from (2.3) by replacing the term $\log[1 - \exp(-\phi_{ij})]$ by the approximation $\log\phi_{ij} - c_{ij}\phi_{ij}$ for each i and j. Good approximations for various neighborhoods of θ_{ij} may be made by taking the value of c_{ij} between 0 and .5. Values of c_{ij} close to .5 work well for neighborhoods of θ_{ij} near 0, whereas smaller values of c_{ij} are appropriate for neighborhoods of large ϕ_{ij}. The approach taken in this section is to determine ways of choosing the c_{ij} so that $\ell_c(\theta)$ will be a good approximation of $\ell(\theta)$.

To concentrate on the error of approximation write $\Delta_c(\theta) = \ell(\theta) - \ell_c(\theta)$, where $\Delta_c(\theta) = \sum_{i=1}^{m}\sum_{j=1}^{k} r_{ij}\Delta_{ij}(\phi_{ij})$ is the total error, and $\Delta_{ij}(\phi_{ij}) = \log[1 - \exp(-\theta_{ij})] - [\log\theta_{ij} - c_{ij}\theta_{ij}]$ is the error for each individual contribution of a failure during the period j with treatment i. For likelihood inference interest is focused on the position of the maximum and the shape of the likelihood function. If $\ell_c(\theta)$ is to be a good approximation to $\ell(\theta)$ for statistical purposes, it is then most desirable that first and second partial derivatives of $\ell(\theta)$ and $\ell_c(\theta)$ be close in neighborhoods of interest. This can be achieved if $\Delta_c(\theta)$ can be made nearly constant in appropriate neighborhoods of θ. However, this condition is rather strong and can only be well approximated when each θ_{ij} is small and each c_{ij} is chosen near .5.

Consider, then, the weaker restriction of selecting c so that the maximum likelihood estimate using $\ell_c(\theta)$, say $\hat{\theta}$, is close to $\tilde{\theta}$, the maximum likelihood estimate using $\ell(\theta)$. Both likelihood functions are well-behaved and the maximizing values may be found by setting partial derivatives equal to 0. Taking derivatives it follows that

$$\dot{\Delta}_c(\theta) = \dot{\ell}(\theta) - \dot{\ell}_c(\theta).$$

Note that if $\dot{\Delta}_c(\tilde{\theta}) = 0$, then since $\dot{\ell}(\tilde{\theta}) = 0$ it follows that $\dot{\ell}_c(\tilde{\theta}) = 0$, and thus $\tilde{\theta} = \hat{\theta}$. So if $\dot{\Delta}_c(\theta)$ can be made approximately 0 near $\tilde{\theta}$, then $\hat{\theta}$ will be close to $\tilde{\theta}$.

For a fixed value θ^o, c can be selected so that $\dot{\Delta}_c(\theta^o) = 0$. Let ϕ be the mk dimensional vector of the ϕ_{ij}. Note

$$\frac{\partial \Delta_c(\theta)}{\partial \theta} = \left[\frac{\partial \phi}{\partial \theta}\right]\left[r_{11}\frac{\partial \Delta_{11}(\phi_{11})}{\partial \phi_{11}}, \ldots, r_{mk}\frac{\partial \Delta_{mk}(\phi_{mk})}{\partial \phi_{mk}}\right]'.$$

Then selecting each c_{ij} so that $\partial \Delta_{ij}(\phi_{ij}^o)/\partial \phi_{ij} = 0$ will make $\partial \Delta_c(\theta^o)/\partial \theta = 0$. This gives

$$\frac{\partial \Delta_{ij}(\phi_{ij}^o)}{\partial \phi_{ij}} = \frac{\exp(-\phi_{ij}^o)}{1 - \exp(-\phi_{ij}^o)} - \frac{1}{\phi_{ij}^o} - c_{ij} = 0,$$

so choosing

$$c_{ij} = \frac{\exp(-\phi_{ij}^o)}{1 - \exp(-\phi_{ij}^o)} - \frac{1}{\phi_{ij}^o} \tag{5.1}$$

makes $\dot{\Delta}_c(\theta^o) = 0$.

This suggests a way to select the c_{ij} to obtain good estimates of θ. Obtain a preliminary estimate of θ, say θ^o, then choose each c_{ij} by

equation (5.1). If $\theta^o \doteq \tilde{\theta}$, then $\hat{\theta} \doteq \tilde{\theta}$, so that for θ reasonably near θ^o, $\hat{\theta}$ can be expected to be a good estimate of θ.

Writing $q_{ij}^o = \exp(-\phi_{ij}^o)$, the problem can be recast in terms of selecting preliminary estimates of q_{ij}, so

$$c_{ij} = -\frac{q_{ij}^o}{1 - q_{ij}^o} - \frac{1}{\log q_{ij}^o},$$

A simple procedure is to use $q_{ij}^o = \hat{q}_{ij} = s_{ij}/(r_{ij} + s_{ij})$. This then leads to selecting

$$c_{ij} = -\frac{1}{\log \hat{q}_{ij}} - \frac{\hat{q}_{ij}}{1 - \hat{q}_{ij}},$$

which is exactly formula (3.3) for choosing $\hat{c}_{ij}$. So the use of $\hat{c}_{ij}$ can be justified on the grounds of attempting to select c so that $\ell(\theta)$ and $\ell_c(\theta)$ will give approximately the same maximum likelihood estimates.

It should be pointed out that $\tilde{\theta}$, the maximum likelihood estimate using $\ell(\theta)$ may be obtained by an iterative procedure employing a sequence $\{\ell_{c_1}(\theta), \ell_{c_2}(\theta), \ldots\}$ of approximate likelihoods. A procedure can be followed which is a direct application of the EM Algorithm as given by Dempster, Laird, and Rubin (1977). The likelihood $\ell(\theta)$ can be viewed as an incomplete data likelihood where exact failure times are unknown, whereas $\ell_c(\theta)$ is a complete data likelihood gotten from exact time data.

The EM Algorithm proceeds as follows:

(1) A preliminary estimate θ^1 is made to start the iteration.

(2) Expectation step: using θ^n and the observed incomplete data find the expected failure times, i.e., find c^n by

$$c_{ij}^{n} = \frac{\exp(-\phi_{ij}^{n})}{1 - \exp(-\phi_{ij}^{n})} - \frac{1}{\phi_{ij}^{n}} .$$

(3) Maximization step: find θ^{n+1} that maximizes $\ell_{c^n}(\theta)$, the complete data likelihood.

(4) Iterate until $\theta^{n} \to \tilde{\theta}$.

Dempster et al. have proved the convergence of θ^{n} to $\tilde{\theta}$ under rather general conditions. Using this procedure one could obtain by iteration $\ell_{c^o}(\theta)$ such that $\ell_{c^n}(\theta) \to \ell_{c^o}(\theta)$, and the maximum likelihood estimate using $\ell_{c^o}(\theta)$ would be $\tilde{\theta}$. Besides having the true maximum likelihood estimates, one would also have the advantages for analysis provided by $\hat{\lambda}(\beta)$ and $\ell^*(\beta)$ with $\ell_{c^o}(\theta)$.

Generally, though, it is felt that this iteration procedure does not gain enough over the previously suggested $\ell_{\hat{c}}(\theta)$ to justify the added computational work. Also for testing hypotheses which require different restrictions on θ, $\ell_{c^o}(\theta)$ will not give the true restricted maximum likelihood estimates. However, if one wishes to work directly with $\ell(\theta)$ following Prentice and Gloeckler (1978), this iterative scheme does provide an alternative to the usual Newton-Raphson method for obtaining $\tilde{\theta}$ and appears to be simpler in many cases.

II.6. Further Examination of the Approximation

An appealing property for the proposed approximation (3.4) is the consistency of the estimator $\hat{\theta}$ found by maximizing $\ell_{\hat{c}}(\theta)$. Previous approximate methods suggested by Cox (1972), Peto (1972b), Breslow (1974), and Holford (1976) fail to give consistent estimators of the parameters.

This has been one of the strongest criticisms leveled against these approximate methods, e.g., Prentice and Gloeckler (1978).

The definition of consistency used here is a generalization of Fisher consistency as given by Rao (1973, p. 345), adapted to this setting in which some of the observations are censored and the data come from m different distributions. For the problem at hand estimators $\hat{\beta}$ and $\hat{\lambda}$ will be called consistent for β and λ, if whenever the observed conditional frequencies $\hat{q}_{ij} = s_{ij}/(r_{ij} + s_{ij})$ are equal to the true conditional probabilities $q_{ij} = \exp\{-\lambda_j \exp(\beta' x_{ij})\}$ for all i and j, then $\hat{\beta} = \beta$ and $\hat{\lambda} = \lambda$. Intuitively the idea is that the estimators are consistent if observing data that exactly fit the theoretical model leads to estimates which are exactly the true parameter values. As with Fisher consistency, strong and weak convergence are implied by this definition when suitable regularity conditions are met. The grouped data setting requires a restriction on censoring to get strong or weak consistency. In order that $\hat{q}_{ij} \to q_{ij}$, one must assume censoring takes place at the endpoints of the time periods. This assumption is needed by Prentice and Gloeckler (1978) to show consistency for the true MLE's.

It is claimed, then, that the estimators $\hat{\beta}$ and $\hat{\lambda}$, found by maximizing $\ell_{\hat{c}}(\beta,\lambda)$ are consistent. To see this, suppose $\hat{q}_{ij} = q_{ij}$ for all i and j. Note that

$$s_{ij} + \hat{c}_{ij} r_{ij} = s_{ij} + \left(-\frac{1}{\log \hat{q}_{ij}} - \frac{\hat{q}_{ij}}{1 - \hat{q}_{ij}}\right) r_{ij}$$

$$= s_{ij} + \left(\frac{1}{\lambda_j \exp(\beta' x_{ij})} - \frac{s_{ij}}{r_{ij}}\right) r_{ij}$$

$$= r_{ij}/\lambda_j \exp(\beta' x_{ij}).$$

Now evaluating $\dot{\ell}^*(\beta)$ at the true value of β we obtain

$$\dot{\ell}^*(\beta) = \sum_{j=1}^{k}\left[\sum_{i=1}^{m} r_{ij}[x_{ij}] - \frac{r_{\cdot j}\sum_{i=1}^{m}(s_{ij}+r_{ij}\hat{c}_{ij})\exp(\beta' x_{ij})[x_{ij}]}{\sum_{i=1}^{m}(s_{ij}+\hat{c}_{ij}r_{ij})\exp(\beta' x_{ij})}\right]$$

$$= \sum_{j=1}^{k}\left\{\sum_{i=1}^{m} r_{ij}[x_{ij}] - \frac{r_{\cdot j}\sum_{i=1}^{m}(r_{ij}/\lambda_j)[x_{ij}]}{\sum_{i=1}^{m}(r_{ij}/\lambda_j)}\right\}$$

$$= \sum_{j=1}^{k}\left\{\sum_{i=1}^{m} r_{ij}[x_{ij}] - \sum_{i=1}^{m} r_{ij}[x_{ij}]\right\}$$

$$= 0.$$

Since $\hat{\beta}$ is the unique value that solves $\dot{\ell}^*(\hat{\beta}) = 0$, it follows that $\hat{\beta} = \beta$. Finally, note that

$$\hat{\lambda}_j = \hat{\lambda}_j(\hat{\beta}) = \hat{\lambda}_j(\beta)$$

$$= r_{\cdot j}/\sum_{i=1}^{m}(s_{ij} + \hat{c}_{ij}r_{ij})\exp(\beta' x_{ij})$$

$$= r_{\cdot j}/\sum_{i=1}^{m}\left[\frac{r_{ij}}{\lambda_j \exp(\beta' x_{ij})}\right]\exp(\beta' x_{ij})$$

$$= \lambda_j,$$

so $\hat{\lambda} = \lambda$, also.

For large sample problems the consistency of $\hat{\theta}$ assures us that $\hat{\theta}$ will be a reasonable estimator. Generally $\hat{\theta}$ is found to be very close to $\tilde{\theta}$, with significant differences occurring only when a chosen model does not fit the data well. One may well ask, then, what has become of

the error introduced by approximating $\ell(\theta)$ by $\ell_{\hat{c}}(\theta)$ if the estimator $\hat{\theta}$ is so close to $\tilde{\theta}$?

The approximation as viewed in section 5 was designed to control $\dot{\ell}_c(\theta)$, making $\dot{\ell}_{\hat{c}}(\theta) \doteq \dot{\ell}(\theta)$ in the region of interest. However, no attempt was made to get $\ddot{\ell}(\theta)$ and $\ddot{\ell}_{\hat{c}}(\theta)$ to agree, and it is here that the approximation has some effect on the analysis. From the perspective of section 3, $\ell_{\hat{c}}(\theta)$ was obtained by assuming the added information of exact failure times and piecewise constant hazard functions. One would expect that $I_{\hat{c}}(\theta) = E[-\ddot{\ell}_{\hat{c}}(\theta)]$, the Fisher information obtained from the approximation, would be greater, i.e., would have larger components, than $I(\theta) = E[-\ddot{\ell}(\theta)]$, the Fisher information obtained from the true likelihood. This indeed is the case, but the gain in information is usually slight.

To investigate this further, the following notation will be used. $\phi = \phi(\theta)$ is written to emphasize that ϕ is obtained as a function of θ by the relation $\phi_{ij} = \lambda_j \exp(\beta' x_{ij})$. Also $\ell(\phi)$ and $\ell_{\hat{c}}(\phi)$ represent the functions of ϕ such that $\ell[\phi(\theta)] = \ell(\theta)$ and $\ell_{\hat{c}}[\phi(\theta)] = \ell_{\hat{c}}(\theta)$, i.e.,

$$\ell(\phi) = \sum_{i=1}^{m} \sum_{j=1}^{k} [r_{ij} \log\{1 - \exp(-\phi_{ij})\} - s_{ij}\phi_{ij}],$$

and

$$\ell_{\hat{c}}(\phi) = \sum_{i=1}^{m} \sum_{j=1}^{k} [r_{ij} \log\phi_{ij} - (s_{ij} + \hat{c}_{ij} r_{ij})\phi_{ij}].$$

Applying the chain rule for partial differentiation, the following expressions are obtained.

$$\ddot{\ell}(\theta) = \left[\frac{\partial\phi(\theta)}{\partial\theta}\right]\left[\frac{\partial^2\ell(\phi)}{\partial\phi^2}\right]\left[\frac{\partial\phi(\theta)}{\partial\theta}\right]' + \sum_{i,j}\left[\frac{\partial^2\phi_{ij}(\theta)}{\partial\theta^2}\right]\left(\frac{\partial\ell(\phi)}{\partial\phi_{ij}}\right)$$

$$\ddot{\ell}_{\hat{c}}(\theta) = \left[\frac{\partial\phi(\theta)}{\partial\theta}\right]\left[\frac{\partial^2\ell_{\hat{c}}(\phi)}{\partial\phi^2}\right]\left[\frac{\partial\phi(\theta)}{\partial\theta}\right]' + \sum_{i,j}\left[\frac{\partial^2\phi_{ij}(\theta)}{\partial\theta^2}\right]\left(\frac{\partial\ell_{\hat{c}}(\phi)}{\partial\phi_{ij}}\right).$$

Now taking expectation with respect to the true grouped data model, and noting for all i and j $E[\partial\ell(\phi)/\partial\phi_{ij}] = 0$, we obtain

$$I(\theta) = E[-\ddot{\ell}(\theta)]$$

$$= \left[\frac{\partial\phi(\theta)}{\partial\theta}\right] E\left[\frac{-\partial^2\ell(\phi)}{\partial\phi^2}\right]\left[\frac{\partial\phi(\theta)}{\partial\theta}\right]'.$$

Similarly, taking expectation with respect to the approximating continuous model and noting again that for all i and j $E[\partial\ell_c(\phi)/\partial\phi_{ij}] = 0$, we have

$$I_{\hat{c}}(\theta) = E[-\ell_{\hat{c}}(\theta)]$$

$$= \left[\frac{\partial\phi(\theta)}{\partial\theta}\right] E\left[\frac{-\partial^2\ell_{\hat{c}}(\phi)}{\partial\phi^2}\right]\left[\frac{\partial\phi(\theta)}{\partial\theta}\right]'.$$

It is seen from these expressions that the error introduced by using $I_{\hat{c}}(\theta)$ instead of $I(\theta)$ comes from the difference in $[-\partial^2\ell(\phi)/\partial\phi^2]$ and $[-\partial^2\ell_{\hat{c}}(\phi)/\partial\phi^2]$. The former is a diagonal matrix with entries $r_{ij}\exp(-\phi_{ij})/[1 - \exp(-\phi_{ij})]^2$, whereas the latter is a diagonal matrix with entries r_{ij}/ϕ_{ij}^2. Comparing a particular entry of $I(\theta)$ to the corresponding entry of $I_{\hat{c}}(\theta)$ amounts to comparing a certain linear combination of the terms $\exp(-\phi_{ij})/[1 - \exp(-\phi_{ij})]^2$ with the same linear combination of ther terms $1/\phi_{ij}^2$.

The error introduced by the approximation is perhaps best characterized as the ratio

$$\rho_{ij} = \frac{(1/\phi_{ij}^2)}{\{\exp(-\phi_{ij})/[1 - \exp(-\phi_{ij})]^2\}},$$

which gives the proportional information gained in cell i,j by using the approximation. To get a notion of how the approximate information compares to the true information, consider the following table (Table 5).

Table 5. Proportional gain of information

q_{ij}	ϕ_{ij}	ρ_{ij}
.9	.105	1.001
.8	.223	1.004
.7	.357	1.011
.6	.511	1.022
.5	.693	1.041
.4	.916	1.072
.3	1.204	1.127

Consider the situation where $q_{ij} = q$ is the same for all i and j. Then ρ_{ij} is the same value, say ρ, for each cell, and it is seen that $I_{\hat{C}}(\theta) = \rho I(\theta)$. Thus asymptotic chi-squared statistics for testing $\beta = 0$ using $\ell_{\hat{C}}(\theta)$ will be too big by approximately a factor of ρ, and confidence regions based on $\ell_{\hat{C}}(\theta)$ will be too small by approximately a factor of $1/\sqrt{\rho}$. Generally, though, examination of the table reveals that ρ is near 1 for most reasonable values of q that one could expect to find in a well designed experiment. For example, with q = .5, chi-square test statistics will be inflated by only a factor of 1.041, whereas confidence regions would be about .980 of their true size. This example would represent a poorly designed experiment, since most of the

experimental units would fail in the early time periods, and later time periods would provide little additional information.

Of course, rarely would an experiment have all the q_{ij} equal. To see the effect of the approximation on the analysis one must look closely at the linear combinations that make up each entry of $I_{\hat{C}}(\theta)$. Generally, a toxicology experiment results in only a few low q_{ij} with the vast majority of the q_{ij} at high values. The larger error introduced by the few low values is usually damped out by the more numerous smaller errors involved in the linear combinations making up the entries in $I_{\hat{C}}(\theta)$. Typically the error involved in the asymptotic theory for convergence to normality is greater than the error of approximation. Thus it is felt that the analysis presented here provides, for most practical grouped survival data problems, methods extremely close to the usual maximum likelihood analysis with the true likelihood, but carries the advantages of eliminating the nuisance parameters λ with $\ell^*(\beta)$.

III. EFFICIENCY

III.1. General Approach

The nature of a nonparametric analysis is that it can be applied to a wide class of problems without making restrictive assumptions about the form of the distributions involved. However, to achieve this generality, nonparametric testing procedures must have rather moderate power properties for all distributions, and so for particular parametric alternatives will be less efficient than a corresponding parametric procedure. The usefulness of a nonparametric analysis may be evaluated by comparing its efficiency to parametric methods for parametric alternatives which one can expect to encounter in practice. If a nonparametric procedure for a problem has high efficiency against parametric alternatives which are often employed for the same problem, then the added scope of application for the nonparametric analysis seems desirable in comparison with a small loss of efficiency for the cases where the parametric models hold true.

In this light the efficiency of the Cox analysis, with its nonparametric $\lambda_o(t)$, against exponential and Weibull models, where $\lambda_o(t)$ is given a parametric form, is an important problem. Preliminary investigation of this problem by Kalbfleisch (1974) and Kalbfleisch and McIntosh (1977) indicates that the Cox analysis has very high efficiency for these parametric models. This is further verified by a simulation study of Lee, Desu, and Gehan (1975). If this is so, then the Cox analysis is to be preferred over a parametric analysis, since the Cox model

will fit a much wider class of distributions while still giving efficient analysis for exponential and Weibull alternatives.

Theoretical investigations concerning the efficiency of the Cox analysis for continuous-time data by Efron (1977) and Oakes (1977) have been highly technical and are difficult to fathom. The efficiency studies in continuous-time settings are made extremely difficult by the indefinite quality of the totally unparameterized $\lambda_o(t)$ in the Cox model. However, this difficulty is completely circumvented by considering the efficiency problem in the grouped data or discrete-time model. The distribution-free analysis here does allow for a general parameterization of $\lambda_o(t)$ by assigning a parameter λ_j to each time period. Exponential and Weibull models simply amount to placing smoothing restrictions on $\lambda_o(t)$, so that the efficiency of the distribution-free analysis compared to these parametric models can be investigated directly through certain restrictions on the λ_j. But before addressing these grouped data efficiency problems, a measure of efficiency will be required.

Let us consider an estimation problem for a real-valued parameter θ. Suppose that T_n is an estimator for θ such that the distribution of T_n is asymptotically $N[\theta, \sigma_T^2(\theta)/n_T]$, i.e., $\sqrt{n_T}(T_n - \theta)/\sigma_T(\theta)$ converges in distribution to a standard normal random variable as the sample size $n_T \to \infty$. Also let S_n be another estimator for θ, such that its distribution is asymptotically $N[\theta, \sigma_S^2(\theta)/n_s]$ for sample size n_s. Assume that $\sigma_T^2(\theta) \leq \sigma_S^2(\theta)$. Then the asymptotic efficiency of using S_n instead of T_n to estimate θ is defined to be

$$(1.1) \qquad e(\theta,S,T) = \frac{\sigma_T^2(\theta)}{\sigma_S^2(\theta)}$$

More explanation of this may be found in Kendall and Stuart (1973, Section 17.28) and in Bickel and Doksum (1977, Section 4.4.C). One interpretation is that it is the asymptotic ratio n_T/n_S of sample sizes required to give T_n and S_n equal variances. Note that since the variances are allowed to depend on θ, that the efficiency depends on the value of θ.

It is also felt that $e(\theta,S,T)$ may be used to define efficiency in an hypothesis testing context. Consider testing $H_o:\theta = 0$ versus $H_a:\theta > 0$ by basing the test on either T_n or S_n. The following heuristic argument is given to motivate the use of $e(\theta,S,T)$ as a measure of efficiency here.

When H_o is true, T_n has an approximate $N[0, \sigma_T^2(0)/n_T]$ distribution while S_n has an approximate $N[0, \sigma_S^2(0)/n_S]$ distribution. To construct size α tests the null hypothesis would be rejected for $T_n > z_\alpha \sigma_T(0)/\sqrt{n_T}$ or for $S_n > z_\alpha \sigma_S(0)/\sqrt{n_S}$, where z_α is the value such that for a standard normal random variable Z, $P(Z > z_\alpha) = \alpha$. Now consider a fixed value $\theta > 0$, and let us calculate the approximate power, say β, of the two tests for this alternative θ value. For T,

$$\beta = P[T > z_\alpha \sigma_T(0)/\sqrt{n_T}]$$

$$= P\left[\frac{\sqrt{n_T}(T-\theta)}{\sigma_T(\theta)} > \frac{\sqrt{n_T}\left(\frac{z_\alpha \sigma_T(0)}{\sqrt{n_T}} - \theta\right)}{\sigma_T(\theta)}\right]$$

$$\doteq P\left[Z > \frac{z_\alpha \sigma_T(0)}{\sigma_T(\theta)} - \frac{\sqrt{n_T}\,\theta}{\sigma_T(\theta)}\right].$$

Suppose we fix β and ask what value of n_T will give an approximate power of β for a size α test? Let z_β be the value such that $P(Z > z_\beta) = \beta$, then set

$$z_\beta = \frac{z_\alpha \sigma_T(0)}{\sigma_T(\theta)} - \frac{\sqrt{n_T}\,\theta}{\sigma_T(\theta)} .$$

Solving for n_T, we obtain

$$n_T(\beta,\alpha,\theta) = \left[\frac{z_\alpha \sigma_T(0)}{\sigma_T(\theta)} - z_\beta\right]^2 \frac{\sigma_T^2(\theta)}{\theta^2}$$

as the sample size required to give an approximate power of β for an alternative θ with a size α test. Similarly for tests based on S,

$$n_S(\beta,\alpha,\theta) = \left[\frac{z_\alpha \sigma_S(0)}{\sigma_S(\theta)} - z_\beta\right]^2 \frac{\sigma_S^2(\theta)}{\theta^2}$$

is the sample size required to give an approximate power of β for an alternative θ with a size α test.

The efficiency of using S instead of T can now be defined as the ratio of sample sizes needed to obtain the power β with the given value θ for size α tests, i.e.,

$$e(\beta,\alpha,\theta,S,T) = \frac{n_T(\beta,\alpha,\theta)}{n_S(\beta,\alpha,\theta)} .$$

To remove the dependence on β and α let $\beta \to 1$. This seems reasonable since high power is desirable and often possible in large sample problems. So define the asymptotic efficiency as

$$e(\theta,S,T) = \lim_{\beta\to 1} e(\beta,\alpha,\theta,S,T)$$

$$= \lim_{\beta\to 1} \frac{n_T(\beta,\alpha,\theta)}{n_S(\beta,\alpha,\theta)}$$

$$= \lim_{z_\beta \to -\infty} \left[\frac{\left\{ \frac{z_\alpha \sigma_T(0)}{\sigma_T(\theta)} - z_\beta \right\}^2 \frac{\sigma_T^2(\theta)}{\theta^2}}{\left\{ \frac{z_\alpha \sigma_S(0)}{\sigma_S(\theta)} - z_\beta \right\}^2 \frac{\sigma_S^2(\theta)}{\theta^2}} \right]$$

$$= \frac{\sigma_T^2(\theta)}{\sigma_S^2(\theta)} .$$

So $e(\theta,S,T)$ also has an interpretation in hypothesis testing as the limit of the ratio of sample sizes required for equal power as the power increases to 1. As defined the asymptotic efficiency depends on the value of θ and will generally vary as θ varies. Often for hypothesis testing problems, asymptotic relative efficiency is defined by letting $\theta \to 0$, thus removing the dependence on θ, e.g., Kendall and Stuart (1973, Section 25.5). This reduces the whole measure of efficiency to a single number appropriate for local alternatives; however, it seems to be too great a reduction in many cases where alternatives quite different from 0 are of interest. Allowing asymptotic efficiency to depend on the alternative values of θ does not seem to cause much difficulty in interpretation and gives more scope for the application of the measure. For the purposes of the work presented here, asymptotic efficiency is then defined by (1.1). One may take either an estimation point of view or an hypothesis testing perspective with this definition.

Returning to the grouped data efficiency problem, two assumptions will be made about the experiment in order to simplify efficiency calculations. First it is assumed that the sample sizes n_i allocated to each treatment combination are the same, say n. This requirement somewhat simplifies asymptotic results, and since many survival experiments are

carried out with equal sample sizes, it seems to be a practical restriction. Second it is assumed that there is no internal censoring, that is, no censoring within the duration of the experiment. Censoring by the termination of the experiment after k time periods will still be considered. Without this restriction one must specify the exact censoring mechanism in order to carry out efficiency calculations. Since it has been assumed throughout this work that nothing is known about censoring, it seems best to simply look at the case of no internal censoring. Also many practical experiments, such as the previous toxicology example, involve no internal censoring. Furthermore, most survival experiments have only low numbers of internally censored observations, so that the efficiency results obtained under the assumption of no such censoring will closely approximate the exact efficiency calculations for these experiments.

Consider now an inference problem for grouped data in which a Cox model is employed with parameters $\beta = (\beta_1, \ldots, \beta_p)'$ and $\alpha = (\alpha_1, \ldots, \alpha_k)'$, where $\alpha_j = \log \lambda_j$ for $j = 1, \ldots, k$. In what follows the parameter α will be considered instead of λ due to the simplicity of the formulas obtained. The model is then written as $q_{ij} = \exp\{-\exp(\alpha_j + \beta' x_{ij})\}$, for covariables x_{ij}.

Concentrating on inference about a particular component of β, say β_1, large sample inference is based on $\hat{\beta}_1$, the maximum likelihood estimator, which is asymptotically normal with mean β_1, and variance $I^{\beta_1\beta_1}(\beta,\alpha)$, the upper left entry of the inverse of the information matrix $I(\beta,\alpha)$. Letting $\eta = (\beta_2, \ldots, \beta_p, \alpha_1, \ldots, \alpha_k)'$, the approximate $\mathrm{Var}(\hat{\beta}_1)$ may be obtained as

$$I^{\beta_1\beta_1}(\beta,\alpha) = [I_{\beta_1\beta_1}(\beta,\alpha) - I_{\beta_1\eta}(\beta,\alpha)\ I_{\eta\eta}^{-1}(\beta,\alpha)\ I_{\eta\beta_1}(\beta,\alpha)]^{-1}.$$

For efficiency calculations it is convenient to work with $\bar{I}(\beta,\alpha) = \frac{1}{n} I(\beta,\alpha)$, so we may write

$$\text{Var}(\hat{\beta}_1) \doteq \frac{1}{n}[\bar{I}_{\beta_1\beta_1}(\beta,\alpha) - \bar{I}_{\beta_1\eta}(\beta,\alpha)\ \bar{I}_{\eta\eta}^{-1}(\beta,\alpha)\ \bar{I}_{\eta\beta_1}(\beta,\alpha)]^{-1}.$$

Suppose we have in mind using a parametric form of $\lambda_o(t)$ instead of leaving it unconstrained. This will amount to some smoothing restriction on the λ_j, or equivalently, smoothing the α_j. For instance, if an exponential modeling is used, $\lambda_o(t)$ is constant and so the α_j would all be restricted to be equal when the time periods are equally spaced. Consider smoothing restrictions on the α_j given by the general form $\alpha = A\gamma$, where A is a $k \times a$ real-valued matrix and γ is an a-dimensional vector of parameters with $a < k$. The same regression expression will be used; however, the underlying hazard has been restricted. Letting A_j be row j of A, the model can now be written as $q_{ij} = \exp\{-\exp(A_j\gamma + \beta' x_{ij})\}$.

Large sample inference about β_1 for this parametric model would be based on the maximum likelihood estimator $\tilde{\beta}_1$. Its distribution would be approximately normal with mean β_1 and variance $I^{\beta_1\beta_1}(\beta,\gamma)$. Again writing $\bar{I}(\beta,\gamma) = \frac{1}{n} I(\beta,\gamma)$ and $\xi = (\beta_2, \ldots, \beta_p, \gamma_1, \ldots, \gamma_a)'$, the asymptotic variance of $\tilde{\beta}_1$ may be expressed as

$$\text{Var}(\tilde{\beta}_1) = \frac{1}{n}[\bar{I}_{\beta_1\beta_1}(\beta,\gamma) - \bar{I}_{\beta_1\xi}(\beta,\gamma)\ \bar{I}_{\xi\xi}^{-1}(\beta,\gamma)\ \bar{I}_{\xi\beta_1}(\beta,\gamma)]^{-1}$$

To measure the efficiency of using the unconstrained Cox model instead of the parametric model with $\alpha = A\gamma$, when in fact the parametric model holds, the efficiency concept of the ratio of asymptotic variances

given in (1.1) may be used. The asymptotic efficiency is then expressed as

$$(1.2) \qquad e(\beta,\gamma,\hat{\beta}_1,\tilde{\beta}_1) = \frac{\bar{I}_{\beta_1\beta_1}(\beta,\alpha) - \bar{I}_{\beta_1\eta}(\beta,\alpha)\ \bar{\bar{I}}^{-1}_{\eta\eta}(\beta,\alpha)\ \bar{I}_{\eta\beta_1}(\beta,\alpha)}{\bar{I}_{\beta_1\beta_1}(\beta,\gamma) - \bar{I}_{\beta_1\xi}(\beta,\gamma)\ \bar{\bar{I}}^{-1}_{\xi\xi}(\beta,\gamma)\ \bar{I}_{\xi\beta_1}(\beta,\gamma)}$$

with the numerator evaluated at $\alpha = A\gamma$, since it is assumed the parametric model holds. For a given choice of β and γ, (1.2) may be used to calculate the efficiency, provided the Fisher information matrices for the two models can be evaluated.

III.2. Some Simplifications

When there is no internal censoring, the data can be reduced to just the r_{ij}, since the s_{ij} can be determined from the r_{ij} and n. In fact, the data may now be viewed as a realization of an experiment with m independent multinomial distributions, one for each treatment combination. There are $k+1$ cells in each multinomial corresponding to the k time periods plus an extra cell for survival past the termination of the experiment. Let r denote the km dimensional vector of the r_{ij}. To employ the efficiency expression of (1.2) it will be necessary to evaluate $\bar{I}(\beta,\alpha) = \frac{1}{n} E[-\ddot{\ell}(r;\beta,\alpha)]$ within this multinomial setting.

Suppose the experimental units are arbitrarily partitioned into n groups indexed by $h = 1, \ldots, n$, so that for each group there are m experimental units with one unit assigned to each treatment combination. Let r_{ijh} be an indicator of whether the individual of group h assigned to treatment i failed during time period j, and let r_h be the km dimensional vector of r_{ijh}. The following relationships hold:

$$r_{ij} = \sum_{h=1}^{n} r_{ijh},$$

$$\ell(r;\beta,\alpha) = \sum_{h=1}^{n} \ell(r_h;\beta,\alpha),$$

$$\ddot{\ell}(r;\beta,\alpha) = \sum_{h=1}^{n} \ddot{\ell}(r_h;\beta,\alpha).$$

Because the r_h are independent and identically distributed, $\bar{I}(\beta,\alpha) = E[-\ddot{\ell}(r_h;\beta,\alpha)]$. Furthermore, the Strong Law of Large Numbers can be applied, so that

$$-\frac{1}{n}\ddot{\ell}(r;\beta,\alpha) = -\frac{1}{n}\sum_{h=1}^{n} \ddot{\ell}(r_h;\beta,\alpha)$$

$$\xrightarrow{a.s.} E[-\ddot{\ell}(r_h;\beta,\alpha)]$$

$$= \bar{I}(\beta,\alpha),$$

where convergence is in the sense that each entry of the matrix converges almost surely. Thus $\bar{I}(\beta,\alpha)$ can be determined as the stochastic limit of $-\frac{1}{n}\ddot{\ell}(r;\beta,\alpha)$ as $n \to \infty$.

Throughout this efficiency study the approximate likelihood function will be used in place of the true likelihood function, so $\ell(\beta,\alpha)$ will simply denote the approximate likelihood, i.e.,

$$\ell(\beta,\alpha) = \sum_{i=1}^{m}\sum_{j=1}^{k}\{r_{ij}(\alpha_j + \beta'x_{ij}) - (s_{ij} + \hat{c}_{ij}r_{ij})\exp(\alpha_j + \beta'x_{ij})\}.$$

The second derivatives of the likelihood may be expressed as follows.

$$\ddot{\ell}_{\beta\beta}(\beta,\alpha) = -\sum_{i=1}^{m}\sum_{j=1}^{k}(s_{ij} + \hat{c}_{ij}r_{ij})\exp(\alpha_j + \beta'x_{ij})[x_{ij}][x_{ij}]',$$

$$\ddot{\ell}_{\beta\alpha_j}(\beta,\alpha) = -\sum_{i=1}^{m}(s_{ij} + \hat{c}_{ij}r_{ij})\exp(\alpha_j + \beta'x_{ij})[x_{ij}],$$

$$\ddot{\ell}_{\alpha_j\alpha_j}(\beta,\alpha) = -\sum_{i=1}^{m}(s_{ij} + \hat{c}_{ij}r_{ij})\exp(\alpha_j + \beta'x_{ij}),$$

$$\ddot{\ell}_{\alpha_j\alpha_h}(\beta,\alpha) = 0, \text{ for } j \neq h.$$

The terms

$$(s_{ij} + \hat{c}_{ij}r_{ij})\exp(\alpha_j + \beta'x_{ij}) = -(s_{ij} + \hat{c}_{ij}r_{ij})\log q_{ij}$$

appear consistently in these expressions. In order to obtain the stochastic limit of $-\frac{1}{n}\ddot{\ell}(\beta,\alpha)$ the limit of $-\frac{1}{n}(s_{ij} + \hat{c}_{ij}r_{ij})\log q_{ij}$ will be found.

Note that $s_{ij}/(r_{ij} + s_{ij}) \xrightarrow{a.s.} q_{ij}$, so that $1/\log[s_{ij}/(r_{ij} + s_{ij})] \xrightarrow{a.s.} 1/\log q_{ij}$. Let π_{ij} be the unconditional probability of response in the j^{th} period under treatment i, i.e., $\pi_{ij} = (\prod_{\ell<j} q_{i\ell})p_{ij}$. Then $r_{ij}/n \xrightarrow{a.s.} \pi_{ij}$. Also recall $\hat{c}_{ij} = -1/\log[s_{ij}/(r_{ij} + s_{ij})] - s_{ij}/r_{ij}$. Combining these facts we have

$$\begin{aligned}
-\frac{(s_{ij} + \hat{c}_{ij}r_{ij})\log q_{ij}}{n} &= \left[\frac{-s_{ij}}{n} + \left\{\frac{1}{\log[s_{ij}/(r_{ij} + s_{ij})]} + \frac{s_{ij}}{r_{ij}}\right\}\frac{r_{ij}}{n}\right]\log q_{ij} \\
&= \left\{\frac{1}{\log[s_{ij}/(r_{ij} + s_{ij})]}\right\}\left(\frac{r_{ij}}{n}\right)\log q_{ij} \\
&\xrightarrow{a.s.} \left(\frac{1}{\log q_{ij}}\right)\pi_{ij}\log q_{ij} \\
&= \pi_{ij}.
\end{aligned}$$

The following results are obtained from the above limit expression.

$$(2.1.i) \qquad -\frac{1}{n}\ddot{\ell}_{\beta\beta}(\beta,\alpha) \xrightarrow{a.s.} \sum_{i=1}^{m}\sum_{j=1}^{k}\pi_{ij}[x_{ij}][x_{ij}]' = \bar{I}_{\beta\beta}(\beta,\alpha),$$

$$(2.1.ii) \qquad -\frac{1}{n}\ddot{\ell}_{\beta\alpha_j}(\beta,\alpha) \xrightarrow{a.s.} \sum_{i=1}^{m}\pi_{ij}[x_{ij}] = \bar{I}_{\beta\alpha_j}(\beta,\alpha),$$

$$(2.1.iii) \qquad -\frac{1}{n}\ddot{\ell}_{\alpha_j\alpha_j}(\beta,\alpha) \xrightarrow{a.s.} \sum_{i=1}^{m}\pi_{ij} = \bar{I}_{\alpha_j\alpha_j}(\beta,\alpha),$$

$$(2.1.iv) \qquad -\frac{1}{n}\ddot{\ell}_{\alpha_j\alpha_h}(\beta,\alpha) = 0 = \bar{I}_{\alpha_j\alpha_h}(\beta,\alpha), \text{ for } j \neq h.$$

Thus, in terms of the π_{ij}, $\bar{I}(\beta,\alpha)$ has a very simple representation.

It is possible to exploit the relationship between efficient scores and the information matrix to obtain $\bar{I}(\beta,\gamma)$ in terms of $\bar{I}(\beta,\alpha)$. Define $S = (S_{\beta_1}, \ldots, S_{\beta_p}, S_{\alpha_1}, \ldots, S_{\alpha_k})' = n^{-\frac{1}{2}}\partial\ell(r;\beta,\alpha)/\partial(\beta,\alpha)$ as the $p+k$ dimensional vector of partial derivatives of $\ell(r;\beta,\alpha)/\sqrt{n}$ with respect to the parameters β and α. It is a standard result that under regularity conditions, such as those imposed by the multinomial problem at hand, that $E(S) = 0$ and $\mathrm{Var}(S) = \bar{I}(\beta,\alpha)$. Also the Central Limit Theorem can be applied, so that as $n \to \infty$ S converges in distribution to a multivariate normal distribution.

Now imposing the relationship $\alpha = \alpha(\gamma) = A\gamma$ and defining $S_\gamma = \partial\ell(r;\beta,\alpha(\gamma))/\partial\gamma$, we have $S_\gamma = [\partial\alpha(\gamma)/\partial\gamma][\partial\ell(r;\beta,\alpha)/\partial\alpha]$ or $S_\gamma = A'S_\alpha$. The following facts are immediate.

$$\bar{I}_{\gamma\gamma}(\beta,\gamma) = \mathrm{Var}(S_\gamma) = \mathrm{Var}(A'S_\alpha)$$

$$= A'\mathrm{Var}(S_\alpha)A = A'\bar{I}_{\alpha\alpha}(\beta,\alpha)A,$$

$$\bar{I}_{\gamma\beta}(\beta,\gamma) = \mathrm{Cov}(S_\gamma, S_\beta) = \mathrm{Cov}(A'S_\alpha, S_\beta)$$

$$= A'\mathrm{Cov}(S_\alpha, S_\beta) = A'\bar{I}_{\alpha\beta}(\beta,\alpha),$$

$$\bar{I}_{\beta\beta}(\beta,\gamma) = \mathrm{Var}(S_\beta) = \bar{I}_{\beta\beta}(\beta,\alpha).$$

The above expressions hold for $\bar{I}(\beta,\alpha)$ evaluated at $\alpha = A\gamma$. Thus $\bar{I}(\beta,\gamma)$ may be obtained from $\bar{I}(\beta,\alpha)$ and A.

Let I_{p-1} be the (p-1)x(p-1) identity matrix, and let

$$B = \begin{bmatrix} I_{p-1} & 0 \\ 0 & A \end{bmatrix}$$

Returning to the efficiency formula of (1.2), write $\eta = B\xi$ and from the above facts it follows that

$$\bar{I}_{\xi\xi}(\beta,\gamma) = B'\bar{I}_{\eta\eta}(\beta,\alpha)B,$$

and

$$\bar{I}_{\xi\beta_1}(\beta,\gamma) = B'\bar{I}_{\eta\beta_1}(\beta,\alpha).$$

(1.2) may then be rewritten as

$$(2.2) \qquad e(\beta,\gamma,\hat{\beta}_1,\tilde{\beta}_1) = \frac{\bar{I}_{\beta_1\beta_1}(\beta,\alpha) - \bar{I}_{\beta_1\eta}(\beta,\alpha)\, I^{-1}_{\eta\eta}(\beta,\alpha)\, I_{\eta\beta_1}(\beta,\alpha)}{\bar{I}_{\beta_1\beta_1}(\beta,\alpha) - I_{\beta_1\eta}(\beta,\alpha)B[B'\bar{I}_{\eta\eta}(\beta,\alpha)B]^{-1}B'\bar{I}_{\eta\beta_1}(\beta,\alpha)}.$$

This formula combined with the simple representation of $\bar{I}(\beta,\alpha)$ given in (2.1) allows for the ready computation of the efficiency of the nonparametric procedure against a parametric procedure with $\alpha = A\gamma$.

It should be noted that the use of the approximate likelihood is not crucial to the application of formula (2.2). It is possible to carry out the efficiency calculations with the true likelihood; however,

the expression for $\bar{I}(\beta,\alpha)$ is more complicated than the simple result of (2.1). It has previously been shown that the approximation is very good, especially for large values of the q_{ij}, so using the approximation in the efficiency calculations should not cause much error. Furthermore, the error in approximation will cause the information expressions of the numerator and the denominator of (2.2) to both be slightly large, and hence the error will have very little effect on the ratio.

The efficient scores may also be used to give a simple view of the efficiency expression (1.2). Assuming S has a multivariate normal distribution, it follows that

$$\begin{aligned}\mathrm{Var}(S_{\beta_1}|S_\eta) &= \mathrm{Var}(S_{\beta_1}) - \mathrm{Cov}(S_{\beta_1},S_\eta)\ \mathrm{Var}^{-1}(S_\eta)\ \mathrm{Cov}(S_\eta,S_{\beta_1}) \\ &= \bar{I}_{\beta_1\beta_1}(\beta,\alpha) - \bar{I}_{\beta_1\eta}(\beta,\alpha)\ \bar{I}^{-1}_{\eta\eta}(\beta,\alpha)\ \bar{I}_{\eta\beta_1}(\beta,\alpha),\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(S_{\beta_1}|S_\xi) &= \mathrm{Var}(S_{\beta_1}) - \mathrm{Cov}(S_{\beta_1},S_\xi)\ \mathrm{Var}^{-1}(S_\xi)\ \mathrm{Cov}(S_\xi,S_{\beta_1}) \\ &= \bar{I}_{\beta_1\beta_1}(\beta,\gamma) - \bar{I}_{\beta_1\xi}(\beta,\gamma)\ \bar{I}^{-1}_{\xi\xi}(\beta,\gamma)\ \bar{I}_{\xi\beta_1}(\beta,\gamma).\end{aligned}$$

So we may think of (2.2) as

$$e(\beta,\gamma,\hat{\beta}_1,\tilde{\beta}_1) = \frac{\mathrm{Var}(S_{\beta_1}|S_\eta)}{\mathrm{Var}(S_{\beta_1}|S_\xi)}.$$

The efficiency is then a comparison of the variance of S_{β_1} when conditioning on S_η with the variance of S_{β_1} if the conditioning is only on S_ξ.

III.3. The Two-Sample Problem

The simplest and most basic survival problem is the comparison of two different samples. The efficiency of the distribution-free approach compared to smooth modeling of the hazard functions will first be considered in this context. It is assumed that n of 2n experimental units are randomized to treatment 1, while the remaining n units are assigned to treatment 2. The number of failures are to be recorded for k time periods for each sample. Time periods will be considered as equally spaced. This is by no means a vital assumption for carrying out the efficiency calculations, but it does simplify things and is a situation often encountered in actual experiments.

To distinguish the samples a simple indicator covariable will be used, i.e., $x_{1j} = 0$ and $x_{2j} = 1$ for all j. So the model gives $q_{1j} = \exp(-\exp \alpha_j)$ and $q_{2j} = \exp\{-\exp(\alpha_j + \beta\}$. Other parameterizations can be used, but one must take care in comparing the results of differently parameterized models. Generally, if parameter values are chosen for two different models so that they both predict the same values for all q_{ij}, then the efficiency results will be the same at those particular points in the parameter spaces. It is also possible to use time dependent covariables to give richer models for the two-sample problem; however, efficiency interpretations for multi-dimensional β are more difficult.

Consider now the simplest possible smoothing restriction on the underlying hazard function, that it is a constant function $\lambda_o(t) = c$. This amounts to assuming that the two survival distributions follow exponential distributions. With the assumption of equally spaced time periods it is seen that

$$\alpha_j = \log \int_{a_{j-1}}^{a_j} \lambda_o(t)dt = \log\{c(a_j - a_{j-1})\} = \gamma$$

for each j. Then α is restricted by the relation $\alpha = A\gamma$, where A is the $k \times 1$ matrix with each entry equal to 1. To study the efficiency of using the nonparametric analysis for inference about β instead of the appropriate parametric analysis when the underlying hazard is constant, expression (2.2) may be applied.

Formulas (2.1) give particularly simple results for the two-sample problem.

(3.1.i) $$\bar{I}_{\beta\beta}(\beta,\alpha) = \sum_{j=1}^{k} \pi_{2j} = \pi_{2\cdot},$$

(3.1.ii) $$\bar{I}_{\beta\alpha_j}(\beta,\alpha) = \pi_{2j},$$

(3.1.iii) $$\bar{I}_{\alpha_j\alpha_j}(\beta,\alpha) = \sum_{i=1}^{2} \pi_{ij} = \pi_{\cdot j},$$

(3.1.iv) $$\bar{I}_{\alpha_j\alpha_h}(\beta,\alpha) = 0, \text{ for } j \neq h.$$

Using these formulas and setting $B = A$ the asymptotic efficiency expression of (2.2) becomes

$$e(\beta,\gamma,\hat{\beta},\tilde{\beta}) = \frac{\bar{I}_{\beta\beta}(\beta,\alpha) - \bar{I}_{\beta\alpha}(\beta,\alpha)\ \bar{I}_{\alpha\alpha}^{-1}(\beta,\alpha)\ \bar{I}_{\alpha\beta}(\beta,\alpha)}{\bar{I}_{\beta\beta}(\beta,\alpha) - \bar{I}_{\beta\alpha}(\beta,\alpha)\, A[A'\bar{I}_{\alpha\alpha}(\beta,\alpha)A]^{-1}\, A\bar{I}_{\alpha\beta}(\beta,\alpha)}$$

$$= \frac{\pi_{2\cdot} - \sum_{j=1}^{k} (\pi_{2j}^2/\pi_{\cdot j})}{\pi_{2\cdot} - (\pi_{2\cdot}^2/\pi_{\cdot\cdot})}$$

(3.2) $$= \frac{\sum_{j=1}^{k} (\pi_{1j}\ \pi_{2j}/\pi_{\cdot j})}{(\pi_{1\cdot}\ \pi_{2\cdot}/\pi_{\cdot\cdot})},$$

where $\pi_{..} = \sum_{i=1}^{2} \sum_{j=1}^{k} \pi_{ij}$.

Recall that the π_{ij} must be computed from the model for particular choices of β and γ. Letting $q_1 = \exp\{-\exp \gamma\}$ and $q_2 = \exp\{-\exp(\gamma + \beta)\} = q_1^{\exp \beta}$ we can write $\pi_{1j} = q_1^{j-1}(1 - q_1)$ and $\pi_{2j} = q_2^{j-1}(1 - q_2)$. Table 5 gives an indication of the efficiency for various choices of the parameters. For ease of interpretation in the discrete setting the table is given in terms of q_1 and $\exp \beta$ instead of γ and β. Only values of $\exp \beta > 1$ need be considered, since the distribution with the smaller hazard may be selected as distribution 1. Ten time periods were used as this seemed typical of many practical experiments.

Table 6. Asymptotic efficiency for the two-sample problem with constant hazard

		exp β				
	1	1.3	1.6	2	3	5
.95	1	1.000	.998	.995	.981	.934
.90	1	.998	.993	.980	.933	.787
q_1 .85	1	.996	.984	.960	.876	.720
.80	1	.993	.975	.940	.839	.685
.75	1	.989	.948	.925	.819	.671

The efficiency for the distribution-free model is seen to be remarkably good. It has the pleasing property of being completely efficient for $\exp \beta = 1$, i.e., $\beta = 0$, and only slowly losing efficiency as $\exp \beta$ increases. In practice one would not be too concerned about efficiency for testing $H_o: \beta = 0$ when β is large, since it would require only small sample size to distinguish the two distributions no matter which procedure was used. Values of $\exp \beta \geq 3$ represent rather enormous

effects in survival testing. For example, with $q_1 = .90$ and $\exp \beta = 3$, then $q_2 = .729$, and after ten time periods there are about 65 percent failures from distribution 1 compared to about 96 percent failures from distribution 2. Also values of $q_1 < .75$ are not of great interest, since these represent poorly designed experiments in which practically all the failures take place in the first few time periods, and later observations are essentially wasted. So in the interesting region with $q_1 \geq .75$ and $\exp \beta \leq 3$, the efficiency is seen to be always greater than .819 and for many cases to be very near 1.

There is an intriguing relationship between (3.2) and correlation in a $2 \times j$ contingency table. Suppose one restricts attention to only the units which failed, in other words, condition on the event of failure during the experiment. Then the conditional probability of failure during period j under treatment i given failure in some cell is $\phi_{ij} = \pi_{ij}/\pi_{\cdot\cdot}$. Let $\phi_{i\cdot} = \sum_{j=1}^{k} \phi_{ij}$, $\phi_{\cdot j} = \phi_{1j} + \phi_{2j}$, and note $\sum_{i=1}^{2} \sum_{j=1}^{k} \phi_{ij} = 1$. A $2 \times j$ table may be formed with these conditional probabilities.

ϕ_{11}	ϕ_{12}	$\cdots$	ϕ_{1k}	$\phi_{1\cdot}$
ϕ_{21}	ϕ_{22}	$\cdots$	ϕ_{2k}	$\phi_{2\cdot}$
$\phi_{\cdot 1}$	$\phi_{\cdot 2}$	$\cdots$	$\phi_{\cdot k}$	1

A measure of association for such tables is the Pearson coefficient of mean square contingency given by Bishop, Fienberg, and Holland (1975, p. 385) as

$$\Phi^2 = \sum_{j=1}^{k} \sum_{i=1}^{2} \left[\frac{(\phi_{ij} - \phi_{i\cdot}\, \phi_{\cdot j})^2}{\phi_{i\cdot}\, \phi_{\cdot j}} \right].$$

This is a generalization of the squared correlation in 2×2 tables and related to the Pearson Chi-Square statistic for testing independence in $2 \times j$ tables. Expression (3.2) for the two-sample efficiency against constant hazard may be rewritten in terms of the ϕ_{ij} as

$$e(\beta,\gamma,\hat{\beta},\tilde{\beta}) = \sum_{j=1}^{k} \left[\frac{\phi_{1j}\,\phi_{2j}}{\phi_{\cdot j}\,\phi_{1\cdot}\,\phi_{2\cdot}} \right]$$

$e(\beta,\gamma,\hat{\beta},\tilde{\beta})$ and Φ^2 are related by the following identity.

$$\Phi^2 + e(\beta,\gamma,\hat{\beta},\tilde{\beta}) = \sum_{j=1}^{k} \sum_{i=1}^{2} \left[\frac{(\phi_{ij} - \phi_{i\cdot}\phi_{\cdot j})^2}{\phi_{i\cdot}\,\phi_{j\cdot}} \right] + \sum_{j=1}^{k} \left[\frac{\phi_{1j}\,\phi_{2j}}{\phi_{\cdot j}\,\phi_{1\cdot}\,\phi_{2\cdot}} \right]$$

$$= \sum_{j=1}^{k} \left[\frac{(\phi_{1j}-\phi_{1\cdot}\phi_{\cdot j})^2}{\phi_{1\cdot}\,\phi_{\cdot j}} + \frac{(\phi_{2j}-\phi_{2\cdot}\phi_{\cdot j})^2}{\phi_{2\cdot}\,\phi_{\cdot j}} + \frac{\phi_{1j}\,\phi_{2j}}{\phi_{\cdot j}\,\phi_{1\cdot}\,\phi_{2\cdot}} \right]$$

$$= \sum_{j=1}^{k} \left[\frac{\begin{array}{l}(\phi_{2\cdot}\phi_{1j}^2 - 2\phi_{2\cdot}\phi_{1j}\phi_{1\cdot}\phi_{\cdot j} + \phi_{2\cdot}\phi_{1\cdot}^2\phi_{\cdot j}^2 + \phi_{1\cdot}\phi_{2j}^2 \\ \quad - 2\phi_{1\cdot}\phi_{2j}\phi_{2\cdot}\phi_{\cdot j} + \phi_{1\cdot}\phi_{2\cdot}^2\phi_{\cdot j}^2 + \phi_{1j}\phi_{2j})\end{array}}{\phi_{\cdot j}\phi_{1\cdot}\phi_{2\cdot}} \right]$$

$$= \sum_{j=1}^{k} \left[\frac{-2\phi_{1\cdot}\phi_{2\cdot}\phi_{\cdot j}^2 + \phi_{1\cdot}\phi_{2\cdot}\phi_{\cdot j}^2 + \phi_{2\cdot}\phi_{1j}^2 + \phi_{1\cdot}\phi_{2j}^2 + \phi_{1j}\phi_{2j}}{\phi_{\cdot j}\phi_{1\cdot}\phi_{2\cdot}} \right]$$

$$= \sum_{j=1}^{k} \left[\frac{\begin{array}{l}-\phi_{1\cdot}\phi_{2\cdot}\phi_{\cdot j}^2 + \phi_{2\cdot}\phi_{1j}(\phi_{\cdot j}-\phi_{2j}) + \phi_{1\cdot}\phi_{2j}(\phi_{\cdot j}-\phi_{1j}) \\ \quad + \phi_{1j}\phi_{2j}\end{array}}{\phi_{\cdot j}\phi_{1\cdot}\phi_{2\cdot}} \right]$$

$$= \sum_{j=1}^{k} \left[\frac{\begin{array}{l}-\phi_{1\cdot}\phi_{2\cdot}\phi_{\cdot j}^2 + \phi_{2\cdot}\phi_{1j}\phi_{\cdot j} + \phi_{1\cdot}\phi_{2j}\phi_{\cdot j} \\ \quad + \phi_{1j}\phi_{2j}(1-\phi_{2\cdot}-\phi_{1\cdot})\end{array}}{\phi_{\cdot j}\phi_{1\cdot}\phi_{2\cdot}} \right]$$

$$= \sum_{j=1}^{k} (-\phi_{\cdot j}) + \sum_{j=1}^{k} \left(\frac{\phi_{1j}}{\phi_{1\cdot}} \right) + \sum_{j=1}^{k} \left(\frac{\phi_{2j}}{\phi_{2\cdot}} \right)$$

$$= -1 + 1 + 1$$

$$= 1.$$

This establishes that $e(\beta,\gamma,\hat{\beta},\tilde{\beta}) = 1 - \Phi^2$. Thus it is seen that the efficiency is high when the $2 \times j$ table exhibits independence. For a given failure if its treatment assignment and time period of failure are considered to be random, then they must be relatively uncorrelated for the efficiency to be high. One may think of this as saying that knowledge about a unit failing in a given time period should not add much information for determining to which treatment it was assigned. This, of course, corresponds to the efficiency problem under study: whether the added knowledge of constant underlying hazard function will influence the information available for determining treatment effect. The distribution-free analysis will not hurt as long as failure time is not highly correlated with treatment over the duration of the experiment.

It must be kept in mind that the possible values of the ϕ_{ij} are restricted by the modeling with β and γ. The only cases in which the table can be independent with $\Phi^2 = 0$ is for $\beta = 0$, which makes each row of the $2 \times j$ table identical. Furthermore, the modeling forces each row to be decreasing in j with $\phi_{ij} > \phi_{ih}$ for $j < h$. This tends to keep Φ^2 low for reasonable treatment differences, and so the efficiency is usually high. Only for gigantic treatment effects will the table ever exhibit enough correlation to give even moderate values of Φ^2.

Focus will now be directed to smoothing the underlying hazard by linear restrictions of the form $\alpha_j = \gamma_1 + \gamma_2 c_j$, where the c_j are fixed real numbers. If the c_j are increasing in j, then these linear restrictions will restrict the underlying hazard to a class of monotone hazard

functions. So this sort of modeling is appropriate to allow the hazard functions to increase or decrease smoothly. Letting $\gamma = (\gamma_1, \gamma_2)'$ and $A = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ c_1 & c_2 & \cdots & c_k \end{bmatrix}'$ the smoothing restriction is $\alpha = A\gamma$. The two-sample model here may be written as $q_{ij} = \exp\{-\exp(\gamma_1 + \gamma_2 c_j)\}$ and $q_{2j} = \exp\{-\exp(\gamma_1 + \gamma_2 c_j + \beta)\}$. This model is of a form that allows (2.2) to be used to evaluate the efficiency of the distribution-free approach for inference about β when the model holds. Formulas (3.1) still apply for calculating $\bar{I}(\beta,\alpha)$, but the π_{ij} must be calculated from the smooth model; hence the π_{ij} depend only on the choice of β and γ. The asymptotic efficiency is then given by

$$e(\beta,\gamma,\hat{\beta},\tilde{\beta}) = \frac{\bar{I}_{\beta\beta}(\beta,\alpha) - \bar{I}_{\beta\alpha}(\beta,\alpha)\ \bar{I}^{-1}_{\alpha\alpha}(\beta,\alpha)\ \bar{I}_{\alpha\beta}(\beta,\alpha)}{\bar{I}_{\beta\beta}(\beta,\alpha) - \bar{I}_{\beta\alpha}(\beta,\alpha)\ A[A'\bar{I}_{\alpha\alpha}(\beta,\alpha)A]^{-1}\ A'I_{\alpha\beta}(\beta,\alpha)}$$

$$(3.3) \qquad = \frac{\pi_{2\cdot} - \sum_{j=1}^{k} (\pi_{2j}^2/\pi_{\cdot j})}{\pi_{2\cdot} - \left\{ \dfrac{\pi_{2\cdot}^2\, W - 2\pi_{2\cdot}\, UV + \pi_{\cdot\cdot} U^2}{\pi_{\cdot\cdot}\, W - V^2} \right\}}$$

where $U = \sum_{j=1}^{k} c_j \pi_{2j}$, $V = \sum_{j=1}^{k} c_j \pi_{\cdot j}$, $W = \sum_{j=1}^{k} c_j^2 \pi_{\cdot j}$.

A two parameter family of smooth hazard functions often used for continuous data problems is the Weibull family. This can be used to model the underlying hazard function of the Cox continuous model as $\lambda_o(t) = a b^a t^{a-1}$, $a > 0$, $b > 0$. Unfortunately, for the grouped data problem this model does not translate to a linear restriction on the α_j; however, certain approximations can be made which will allow the efficiency expression of (3.3) to be used.

Again assume that the time periods are equally spaced, and for simplicity let time be scaled so that the end points of the time periods are integers, hence $a_j = j$. If the underlying hazard function has Weibull form, then

$$\begin{aligned}
\alpha_j &= \log \int_{j-1}^{j} \lambda_o(t)dt \\
&= \log \int_{j-1}^{j} a\, b^a\, t^{a-1}\, dt \\
&= \log \{b^a[j^a - (j-1)^a]\} \\
&\doteqdot \log \{a\, b^a\, (j - \tfrac{1}{2})^{a-1}\} \\
&= \log\, (a\, b^a) + (a-1)\, \log\, (j - \tfrac{1}{2}).
\end{aligned}$$

The approximation is made by replacing $\lambda_o(t)$ by a step function with the value on each time period held constant to the value of $\lambda_o(t)$ evaluated at the midpoint of the interval. Setting $\gamma_1 = \log\, (a\, b^a)$, $\gamma_2 = a - 1$, and $c_j = \log\, (j - \frac{1}{2})$ the approximate linear restriction can be written as $\alpha_j = \gamma_1 + \gamma_2 c_j$.

To get some idea of how efficient the continuous Cox model analysis is against Weibull alternatives, expression (3.3) will be applied to the group data setting assuming $\alpha_j = \gamma_1 + \gamma_2 c_j$. Actually this discrete data model is of considerable interest in its own right as linear models in log-time are appealing. Electing to use ten time periods, the scale parameter b is of most interest for small values, since large b implies practically all failures will occur in the early time periods. Table 7 shows the results for various selections of a, b, and exp β.

Table 7. Asymptotic efficiency for the two-sample problem with Weibull hazard

		exp β			
		1	2	3	5
	a				
b = 1	.5	1	.999	.998	.974
	1	1	.999	.996	.981
	2	1	.999	.999	.997
b = .2	.5	1	.999	.998	.995
	1	1	.992	.980	.978
	2	1	.970	.938	.895
b = .1	.5	1	.999	.998	.995
	1	1	.997	.988	.964
	2	1	.995	.978	.925

The efficiency for the distribution-free analysis is even better against this grouped data Weibull model than against the constant hazard model. One might predict this by considering the hierarchy of models in which the one-dimensional constant hazard models are contained in the two-dimensional Weibull models which in turn are contained in the k-dimensional distribution-free models. The Weibull models are a step closer to the distribution-free models than are the constant hazard models. This relationship will be further exploited by the geometrical interpretation given in the next section. The added scope of the distribution-free model makes it the much preferred way to test for treatment effects. Of course, if one wishes to estimate the underlying hazard function and suspects that it is monotone, then Weibull modeling may be beneficial to that end. But for purposes of inference about β

there is almost nothing to be gained by using the Weibull analysis instead of the distribution-free approach.

III.4. The Geometry of the Two-Sample Problem

The two-sample efficiency problem can be neatly characterized by considering a normed vector space associated with the efficient score vector $S = (S_\beta, S_{\alpha_1}, \ldots, S_{\alpha_k})'$. Since asymptotic efficiency is under study, S will be taken to have a multivariate normal distribution with mean 0 and variance-covariance matrix given by $\Sigma = \bar{I}(\beta,\alpha)$. Recall

$$\Sigma = \begin{bmatrix} \pi_{2\cdot} & \pi_{21} & \pi_{22} & \pi_{23} & \cdots & \pi_{2k} \\ \pi_{21} & \pi_{\cdot 1} & 0 & 0 & \cdots & 0 \\ \pi_{22} & 0 & \pi_{\cdot 2} & 0 & \cdots & 0 \\ \pi_{23} & 0 & 0 & \cdot & & \cdot \\ \vdots & \vdots & \vdots & & \cdot & \cdot \\ \vdots & \vdots & \vdots & & & \cdot \\ \pi_{2k} & 0 & 0 & \cdot & \cdot \cdot & \pi_{\cdot k} \end{bmatrix}$$

and note that Σ is a positive definite matrix. Consider the normed vector space R^{k+1} endowed with the inner product given by $(x,y) = \mathrm{Cov}(x'S, y'S) = x'\Sigma y$ for $x,y \in R^{k+1}$. The norm of $x \in R^{k+1}$ is then defined by the relation $||x|| = \sqrt{(x,x)}$, so that $||x||^2 = x'\Sigma x = \mathrm{Var}(x'S)$. Each $x \in R^{k+1}$ may be associated with the linear combination $x'S$ of score statistics, and $||x||$ is the standard deviation of $x'S$. For example, associate $b = (1, 0, 0, \ldots, 0)'$ with S_β, since $b'S = S_\beta$ and $||x||^2 = \mathrm{Var}(S_\beta)$.

Let W be a full rank $(k+1) \times h$ matrix, then the range of W, written $\underline{R}(W)$, is a subspace of R^{k+1}. The orthogonal projection operator on

$\underline{R}(W)$ is given by $P_W = W(W'\Sigma W)^{-1}W'\Sigma$. It can be shown that $P_W^2 = P_W$ and $P_W'\Sigma P_W = \Sigma P_W$; see Rao (1973, pp. 46-48). Furthermore, for any $x \in R^{k+1}$, $x = P_W x + (x - P_W x)$ with $(P_W x, x - P_W x) = 0$ and $||x||^2 = ||P_W x||^2 + ||x - P_W x||^2$. x, then, can be written as a sum of two vectors, one in $\underline{R}(W)$ and the other in the space orthogonal to $\underline{R}(W)$. There is a direct connection between conditional variances of the score statistics and these orthogonal projection operators. Let x'S be a linear combination of the scores and W'S be a vector of some other linear combinations of the scores. Then

$$\begin{aligned}
\mathrm{Var}(x'S|W'S) &= \mathrm{Var}(x'S) - \mathrm{Cov}(x'S,W'S)\,\mathrm{Var}^{-1}(W'S)\,\mathrm{Cov}(W'S,x'S) \\
&= x'\Sigma x - x'\Sigma W(W'\Sigma W)^{-1}W'\Sigma x \\
&= x'\Sigma x - x'\Sigma P_W x \\
&= x'\Sigma x - x'P_W'\Sigma P_W x \\
&= ||x||^2 - ||P_W x||^2 \\
&= ||x - P_W x||^2 .
\end{aligned}$$

$\mathrm{Var}(x'S|W'S)$ is seen to be the squared norm of the part of x orthogonal to $\underline{R}(W)$.

From the results of section 2,

$$e(\beta,\gamma,\hat{\beta},\tilde{\beta}) = \frac{\mathrm{Var}(S_\beta|S_\alpha)}{\mathrm{Var}(S_\beta|S_\gamma)} ,$$

where $S_\gamma = A'S_\alpha$ for a model $\alpha = A'\gamma$. Let $C = \begin{bmatrix} 0 \\ I_k \end{bmatrix}$, I_k being the k-dimensional identity matrix, and let $D = \begin{bmatrix} 0 \\ A \end{bmatrix}$. Then $\underline{R}(D) \subset \underline{R}(C)$, $S_\alpha = C'S$, and $S_\gamma = D'S$. It follows that

$$e(\beta,\gamma,\hat{\beta},\tilde{\beta}) = \frac{\mathrm{Var}(b'S|C'S)}{\mathrm{Var}(b'S|D'S)}$$

$$= \frac{||b||^2 - ||P_C b||^2}{||b||^2 - ||P_D b||^2}$$

$$= \frac{||b - P_C b||^2}{||b - P_D b||^2} .$$

Figure 2 helps to visualize this geometric interpretation of the asymptotic efficiency. The picture is a three-dimensional analogue of the $k+1$ dimensional problem and is distorted so that orthogonality appears as the usual right angles familiar to intuition.

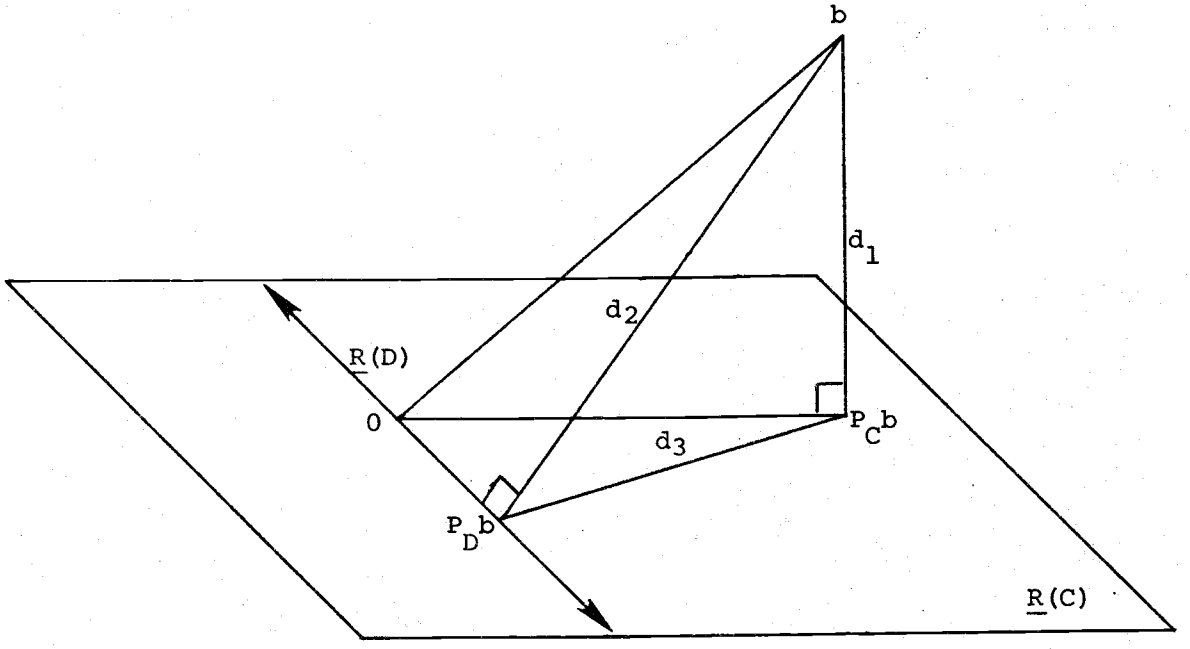

Figure 2. The two-sample geometric picture

The norm of a vector may be interpreted as the length of a vector in this space, so the asymptotic efficiency amounts to comparing the lengths $d_1 = ||b - P_D b||$ and $d_2 = ||b - P_C b||$ by $e(\beta,\gamma,\hat{\beta},\tilde{\beta}) = d_1^2/d_2^2$. Considering the

distance $d_3 = ||P_C b - P_D b||$ and noting $d_1^2 + d_3^2 = d_2^2$, the efficiency may be given as $e(\beta,\gamma,\hat{\beta},\tilde{\beta}) = d_1^2/(d_1^2 + d_3^2)$. The efficiency depends on how close $P_C b$ is to $\underline{R}(D)$ as measured by d_3 in comparison with d_1. Comparing d_3 with d_1 by the ratio $r = d_3^2/d_1^2$, the efficiency may be written as

$$e(\beta,\gamma,\hat{\beta},\tilde{\beta}) = \frac{d_1^2}{d_1^2 + d_3^2}$$

$$= \frac{1}{1 + \frac{d_3^2}{d_1^2}}$$

$$= \frac{1}{1 + r} .$$

The high efficiency of the distribution-free analysis may be explained by considering why $P_C b$ is usually close to $\underline{R}(D)$ when D is formed from a smooth model $\alpha = A\gamma$. Calculating $P_C b$ gives

$$P_C b = C(C'\Sigma C)^{-1} C'\Sigma b$$

$$= \left(0, \frac{\pi_{21}}{\pi_{\cdot 1}}, \frac{\pi_{22}}{\pi_{\cdot 2}}, \ldots, \frac{\pi_{2k}}{\pi_{\cdot k}}\right)' .$$

Since $\underline{R}(D) = \underline{R}\begin{bmatrix} 0 \\ A \end{bmatrix}$ and $P_C b$ are both contained in the k-dimensional subspace $\underline{R}(C)$ formed by restricting all first coordinates to 0, the efficiency problem amounts to measuring how close the k-dimensional vector $c = \left(\frac{\pi_{21}}{\pi_{\cdot 1}}, \frac{\pi_{22}}{\pi_{\cdot 2}}, \ldots, \frac{\pi_{2k}}{\pi_{\cdot k}}\right)'$ is to $\underline{R}(A)$. The k + 1-dimensional space endows this k-dimensional subspace with an inner product given by $(x,y) = x'\Sigma_{\alpha\alpha} y$, where $x,y \in R^k$ and

$$\Sigma_{\alpha\alpha} = \begin{bmatrix} \pi_{\cdot 1} & 0 & 0 & \cdots & 0 \\ 0 & \pi_{\cdot 2} & 0 & \cdots & 0 \\ 0 & 0 & \cdot & & \cdot \\ \vdots & \vdots & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot \cdot & \pi_{\cdot k} \end{bmatrix}.$$

d_3^2, the measure of how close c is to $\underline{R}(A)$, may be calculated within this k-dimensional normed vector space by

$$d_3^2 = ||c - P_A c||^2$$

$$= \sum_{j=1}^{k} \pi_{\cdot j} \left(\frac{\pi_{2j}}{\pi_{\cdot j}} - p_j \right)^2,$$

where $P_A c = A(A'\Sigma_{\alpha\alpha}A)^{-1}A'\Sigma_{\alpha\alpha}c = (p_1, p_2, \ldots, p_k)'$. This corresponds to a weighted least squares problem by the well-known fact that

$$d_3^2 = \min_{x \in \underline{R}(A)} \left\{ \sum_{j=1}^{k} \pi_{\cdot j} \left(\frac{\pi_{2j}}{\pi_{\cdot j}} - x_j \right)^2 \right\},$$

where $x = (x_1, x_2, \ldots, x_k)'$. So the measure of closeness d_3^2 is just the minimal weighted least squares value for the regression of c onto $\underline{R}(A)$, the weights being given by the proportion of units failing in each time period.

The complete asymptotic efficiency at $\beta = 0$ for the constant hazard and Weibull models is now easily explained. For all j, $\pi_{1j} = \pi_{2j}$ when $\beta = 0$, and hence $\pi_{2j}/\pi_{\cdot j} = \frac{1}{2}$, so $c = (\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2})'$. $c \in \underline{R}(A)$ whenever A contains the subspace spanned by $(1, 1, \ldots, 1)'$, thus $d_3^2 = 0$ for $\beta = 0$, and $e(0,\gamma,\hat{\beta},\tilde{\beta}) = d_1^2/(d_1^2 + d_3^2) = 1$. For smooth modeling of

the hazard function by $\alpha = A'\gamma$, one would certainly want to include the possibility of constant hazard with $\alpha_j = \gamma$ for all j, so reasonable smoothing models will always have $(1, 1, \ldots, 1)' \in \underline{R}(A)$. The efficiency for β values near 0 is then expected to be high, but the results of the previous section indicate that the efficiency remains high for even moderately high β values which one would expect to encounter in real problems. Further investigation of the nature of the vector $c = (\frac{\pi_{21}}{\pi_{\cdot 1}}, \frac{\pi_{22}}{\pi_{\cdot 2}}, \ldots, \frac{\pi_{2k}}{\pi_{\cdot k}})'$ will help to see why this is so.

The following inequality will be required. Let a, b, x be real numbers such that $0 < a,b < 1$ and $x \geq 1$, then

$$\frac{1 - b^x}{1 - b} \leq \frac{(1 - a^x)\, a^{1-x}}{1 - a} . \tag{4.1}$$

That this is true can be seen by considering the two differentiable functions $f(x) = (1 - b^x)/1 - b$ and $g(x) = (1 - a)^x a^{1-x}/(1 - a)$ defined on $[1,\infty)$. Note $f(1) = g(1)$, so considering the derivatives, if $\dot{f}(x) < \dot{g}(x)$ can be established for all $x \geq 1$, this will imply $f(x) \leq g(x)$. The derivatives are given by $\dot{f}(x) = -b^x \log b/(1 - b)$ and $\dot{g}(x) = -a^{1-x} \log a/(1 - a)$. Note that for all $a \in (0,1)$ that $1 - a < -\log a$, or $-\log a/(1 - a) > 1$. Also for all $b \in (0,1)$, $(1 - b)/b > -\log b$; so $-b \log b/(1 - b) < 1$. Hence $\dot{f}(1) = -b^x \log b/(1 - b) < 1 < -\log a/(1 - a) = \dot{g}(1)$. Furthermore, $\ddot{f}(x) = -b^x (\log b)^2/(1 - b) < 0$, and $\ddot{g}(x) = a^{1-x}(\log a)^2/(1 - a) > 0$, so since $\dot{f}(1) < \dot{g}(1)$, it follows that $\dot{f}(x) < \dot{g}(x)$ for all $x \geq 1$. Thus inequality (4.1) holds.

Using inequality (4.1) it can be shown that the entries of c are monotone decreasing for any choice of α. More specifically, for $j = 1, \ldots, k - 1$ and $h = j + 1$ it is claimed that $\pi_{2j}/\pi_{\cdot j} \geq \pi_{2h}/\pi_{\cdot h}$. Applying

(4.1) with $a = q_{1j}$, $b = q_{1h}$ and $x = \exp \beta$

$$\frac{1 - q_{1h}^{\exp \beta}}{1 - q_{1h}} \leq \frac{(1 - q_{1j}^{\exp \beta})\, q_{1j}^{1 - \exp \beta}}{1 - q_{1j}} .$$

Recall $q_{2j} = q_{1j}^{\exp \beta}$ and $q_{2h} = q_{1h}^{\exp \beta}$, so

$$\frac{1 - q_{2h}}{1 - q_{1h}} \leq \frac{(1 - q_{2j})\, q_{1j}}{(1 - q_{1j})\, q_{2j}} ,$$

or

$$\frac{q_{2j}(1 - q_{2h})}{1 - q_{2j}} \leq \frac{q_{1j}(1 - q_{1h})}{1 - q_{1j}} . \tag{4.2}$$

Using (4.2) the claim is established, since

$$\begin{aligned}
\frac{\pi_{2j}}{\pi_{\cdot j}} &= \frac{\left(\prod_{\ell=1}^{j-1} q_{2\ell}\right)(1 - q_{2j})}{\left(\prod_{\ell=1}^{j-1} q_{1\ell}\right)(1 - q_{1j}) + \left(\prod_{\ell=1}^{j-1} q_{2\ell}\right)(1 - q_{2j})} \\
&\geq \frac{\left(\prod_{\ell=1}^{j-1} q_{2\ell}\right)(1-q_{2j})\left[\frac{q_{2j}(1-q_{2h})}{1-q_{2j}}\right]}{\left(\prod_{\ell=1}^{j-1} q_{1\ell}\right)(1-q_{1j})\left[\frac{q_{1j}(1-q_{1h})}{1-q_{1j}}\right] + \left(\prod_{\ell=1}^{j-1} q_{2\ell}\right)(1-q_{2j})\left[\frac{q_{2j}(1-q_{2h})}{1-q_{2j}}\right]} \\
&= \frac{\left(\prod_{\ell=1}^{j} q_{2\ell}\right)(1-q_{2h})}{\left(\prod_{\ell=1}^{j} q_{1\ell}\right)(1-q_{1h}) + \left(\prod_{\ell=1}^{j} q_{2\ell}\right)(1-q_{2h})} \\
&= \frac{\pi_{2h}}{\pi_{\cdot h}} .
\end{aligned}$$

The fact that the entries of c are decreasing helps to explain why the efficiency is usually high for smooth modeling of the form $\alpha = A\gamma$. A good smooth model will employ a matrix A which will allow for the possibility of a decreasing hazard function with $\alpha_1 > \alpha_2 > \ldots > \alpha_k$, so $\underline{R}(A)$ will contain a fairly rich set of decreasing vectors. Then the decreasing vector c is apt to be near $\underline{R}(A)$, i.e., $\underline{R}(A)$ will provide a good fit for the vector c as measured by the weighted least squares regression. The entries of c decrease rather smoothly for moderate values of β and reasonably large values of the q_{ij}, so a smooth A will usually fit c well. When β gets really large, $\pi_{21}/\pi_{\cdot 1}$ is near 1, while all the other entries of c are near 0. In this case c is somewhat unsmooth, so that A may not fit c too well, and the efficiency is usually not high.

Of course, it is possible to select a matrix A so that the asymptotic efficiency is not high even for local alternatives to $\beta = 0$, but this forces the model $\alpha = A\gamma$ to give very jagged hazard functions. However, no matter what A is chosen, the efficiency will never be lower than $\frac{1}{2}$ at $\beta = 0$. This is seen by noting $d_1^2 = \frac{1}{2}\pi_{2\cdot}$ when $\beta = 0$, and that the maximum value d_2^2 can ever be is $\pi_{2\cdot}$. It is clear now why one should expect the distribution-free approach to have greater efficiency for a richer model such as the Weibull family than for the constant hazard model. The richer model will span a higher dimensional subspace with $\underline{R}(A)$, and this will generally fit c more closely and give smaller values of d_3^2. The geometric interpretation of the asymptotic efficiency brings out the facts that the distribution-free analysis is especially geared to do well for local alternatives to $\beta = 0$ and for modeling that includes constant and smoothly decreasing hazard functions. These two

properties are very compelling arguments for the application of the distribution-free analysis instead of a parametric analysis when inference about β is of prime concern.

III.5. Other Efficiency Problems

This efficiency investigation has concentrated on the two-sample problem to this point. The two-sample problem brings out the basic properties of the efficiency of the distribution-free approach, and investigation of the efficiency in other factorial designs essentially involves extensions of these ideas. The tools developed in sections 1 and 2 allow the direct calculation of efficiency for other designs. Formulas (2.1) may be used to set up the information matrix $\bar{I}(\beta,\alpha)$, then for inference about a particular component of β, when $\alpha = A\gamma$, efficiency calculation is straightforward using expression (2.2). Further investigation may be carried out by considering the normed vector space associated with the score statistics as was done for the two-sample problem.

Attention will now be directed to a simple 2×2 factorial design. This problem will give some idea of how efficient the distribution-free approach is when there are nuisance parameters in the β vector. Suppose there are two factors A and B to consider, and four treatments are run for the factorial design of presence and absence of each factor. Instead of indexing the treatments by single integers i, the more descriptive double subscript ab will be used, so that factor A is absent for $a = 1$ and present for $a = 2$, while factor B is absent for $b = 1$ and present for $b = 2$. A distribution-free model for this experiment is given by $q_{abj} = \exp\{-\exp(\alpha_j + \beta' x_{ab})\}$, where $\beta = \begin{bmatrix}\beta_1 \\ \beta_2\end{bmatrix}$, $x_{11} = \begin{bmatrix}0 \\ 0\end{bmatrix}$,

$x_{12} = \begin{bmatrix}0\\1\end{bmatrix}$, $x_{21} = \begin{bmatrix}1\\0\end{bmatrix}$, and $x_{22} = \begin{bmatrix}1\\1\end{bmatrix}$. Assume that inference is primarily focused on factor A, so that β_1 is of interest, while $\beta_2, \alpha_1, \alpha_2, \ldots, \alpha_k$ are considered nuisance parameters.

Again let

$$\pi_{abj} = \left(\prod_{\ell=1}^{j-1} q_{ab\ell}\right)(1 - q_{abj}),$$

and write

$$\pi_{a\cdot\cdot} = \sum_{b=1}^{2} \sum_{j=1}^{k} \pi_{abj},$$

$$\pi_{\cdot b\cdot} = \sum_{a=1}^{2} \sum_{j=1}^{k} \pi_{abj},$$

$$\pi_{\cdot\cdot j} = \sum_{a=1}^{2} \sum_{b=1}^{2} \pi_{abj},$$

and

$$\pi_{\cdot\cdot\cdot} = \sum_{a=1}^{2} \sum_{b=1}^{2} \sum_{j=1}^{k} \pi_{abj}.$$

The information matrix $\bar{I}(\beta,\alpha)$ may be obtained from formulas (2.1). The entries are given by

$$\bar{I}_{\beta_1\beta_1}(\beta,\alpha) = \pi_{2\cdot\cdot},$$

$$\bar{I}_{\beta_1\beta_2}(\beta,\alpha) = \pi_{22\cdot},$$

$$\bar{I}_{\beta_2\beta_2}(\beta,\alpha) = \pi_{\cdot 2\cdot},$$

$$\bar{I}_{\beta_1\alpha_j}(\beta,\alpha) = \pi_{2\cdot j},$$

$$\bar{I}_{\beta_2\alpha_j}(\beta,\alpha) = \pi_{\cdot 2j},$$

$$\bar{I}_{\alpha_j\alpha_j}(\beta,\alpha) = \pi_{\cdot\cdot j},$$

and

$$\bar{I}_{\alpha_j\alpha_h}(\beta,\alpha) = 0 \text{ for } j \neq h.$$

Suppose the underlying hazard function is constant, so that $\alpha = A\gamma$ where $A = [1, 1, \ldots, 1]'$ and $\gamma \in R$. For given values of β_1, β_2, and γ expression (2.2) may be applied to calculate the efficiency for using the distribution-free approach for inference about β_1 instead of the appropriate parametric analysis. This asymptotic efficiency may be alternatively expressed as

$$e(\beta,\gamma,\hat{\beta}_1,\tilde{\beta}_1) = \frac{\mathrm{Var}(S_{\beta_1} \mid S_{\beta_2}, S_\alpha)}{\mathrm{Var}(S_{\beta_1} \mid S_{\beta_2}, S_\gamma)}.$$

In this case the efficiency may be calculated by

$$e(\beta,\gamma,\hat{\beta}_1,\tilde{\beta}_1) = \frac{\sum_{j=1}^{k}\left(\frac{\pi_{1\cdot j}\pi_{2\cdot j}}{\pi_{\cdot\cdot j}}\right) - \frac{\left[\sum_{j=1}^{k}\left(\frac{\pi_{11j}\pi_{22j} - \pi_{12j}\pi_{21j}}{\pi_{\cdot\cdot j}}\right)\right]^2}{\sum_{j=1}^{k}\left(\frac{\pi_{\cdot 1j}\pi_{\cdot 2j}}{\pi_{\cdot\cdot j}}\right)}}{\left(\frac{\pi_{1\cdot\cdot}\pi_{2\cdot\cdot}}{\pi_{\cdots}}\right) - \frac{\left(\frac{\pi_{11\cdot}\pi_{22\cdot} - \pi_{12\cdot}\pi_{21\cdot}}{\pi_{\cdot\cdot j}}\right)^2}{\left(\frac{\pi_{\cdot 1\cdot}\pi_{\cdot 2\cdot}}{\pi_{\cdots}}\right)}}.$$

Let $q_{11} = \exp(-\exp\gamma)$, then assuming constant hazard the model may be expressed as $q_{11j} = q_{11}$, $q_{12j} = q_{11}^{\exp\beta_2}$, $q_{21j} = q_{11}^{\exp\beta_1}$, and $q_{22j} = q_{11}^{\exp(\beta_1+\beta_2)}$, for all j. Table 8 shows the efficiencies obtained for an experiment with ten time periods for $q_{11} = .9$ and various selections of $\exp\beta_1$ and $\exp\beta_2$.

Table 8. Asymptotic efficiency for a 2 × 2 factorial experiment (ten time periods, $q_{11} = .9$)

		exp β_1			
		1	2	3	5
exp β_2	1	1	.980	.933	.787
	2	1	.964	.892	.752
	3	1	.955	.882	.751
	5	1	.953	.886	.772

The results are very similar to the two-sample problem; in fact, when $\beta_2 = 0$ (exp $\beta_2 = 1$), the numbers are the same as those obtained in Table 6 for $q_1 = .9$. Apparently, the efficiency against the constant hazard model is somewhat less for values of $\beta_2 > 0$ than for $\beta_2 = 0$, although the differences are slight. The distribution-free approach gives complete asymptotic efficiency at $\beta_1 = 0$, no matter what β_2 might be, and only slowly loses efficiency as β_1 increases. The efficiency must improve for richer modeling of the underlying hazard function, so again the distribution-free approach is highly efficient for smooth parametric alternatives.

One of the strong points for Cox models is the use of time dependent covariables. This provides the flexibility to fit most practical data sets and to analyze time dependent treatment effects. To allow for a rich set of alternatives a model with a time dependent covariable x_{ij} will usually be employed in conjunction with a covariable, say z_i, which is not time dependent. For example, in a two-sample problem $q_{ij} = \exp\{-\exp(\alpha_j + z_i\beta_1 + x_{ij}\beta_2)\}$, $z_1 = 0$, $z_2 = 1$, $x_{1j} = 0$ for all j, and $x_{2j} = j$, is a flexible way to model the difference of treatment 2 from treatment

1. An important question here is whether the distribution-free approach remains highly efficient against smooth modeling of the underlying hazard function when time dependent covariables are used. In this setting both β_1 and β_2 are of interest, and it is not appropriate to consider inference about one while treating the other as a nuisance parameter. A new measure of efficiency is required which will handle more than one parameter of interest. That problem is outside the scope of this thesis; however, a related model with only the time dependent covariable may be considered to get a notion of how time dependent covariables affect the efficiency.

Consider, then, the two-sample problem modeled by $q_{ij} = \exp\{-\exp(\alpha_j + x_{ij}\beta)\}$, where $x_{1j} = 0$ for all j, and x_{2j} is some smooth function of j. This model is not as useful as the previous one, but its efficiency can be analyzed by the methods developed here. Now focus on the smoothing of the underlying hazard by $\alpha = A\gamma$ to address the efficiency of using the distribution-free inference for β in this context. The information matrix here may be expressed as

$$\Sigma = \begin{bmatrix} \sum_{j=1}^{k} x_{2j}^2 \pi_{2j} & x_{21}\pi_{21} & x_{22}\pi_{22} & \cdots & x_{2k}\pi_{2k} \\ x_{21}\pi_{21} & \pi_{\cdot 1} & 0 & \cdots & 0 \\ x_{22}\pi_{22} & 0 & \pi_{\cdot 2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{2k}\pi_{2k} & 0 & 0 & \cdots & \pi_{\cdot k} \end{bmatrix}.$$

Taking the geometric approach to efficiency, let us see how the vector $b = (1, 0, \ldots, 0)'$, (associated with $S_\beta = b'S$) projected onto $\underline{R}\begin{bmatrix}0\\ I_k\end{bmatrix}$ (associated with $S_\alpha = \begin{bmatrix}0\\ I_k\end{bmatrix}'S$) will line up with $\underline{R}\begin{bmatrix}0\\ A\end{bmatrix}$ (associated with $S_\gamma = \begin{bmatrix}0\\ A\end{bmatrix}'S$). Letting $D = \begin{bmatrix}0\\ I_k\end{bmatrix}$, then

$$P_D b = D(D'\Sigma D)^{-1}D'\Sigma b$$

$$= (0, x_{21}\pi_{21}/\pi_{\cdot 1}, x_{22}\pi_{22}/\pi_{\cdot 2}, \ldots, x_{2k}\pi_{2k}/\pi_{\cdot k})'.$$

When $\beta = 0$, it is seen that $\pi_{2j}/\pi_{\cdot j} = \frac{1}{2}$ for all j, and so $P_D b = \frac{1}{2}(0, x_{21}, x_{22}, \ldots, x_{2k})'$. Here at last is a case in which the efficiency will not be 1 at $\beta = 0$ when the underlying hazard function is constant, since for $A = [1, 1, \ldots, 1]'$, $P_D b \notin \underline{R}\begin{bmatrix}0\\ A\end{bmatrix}$. However, if $(x_{21}, x_{22}, \ldots, x_{2k})' \in \underline{R}(A)$, the efficiency will be 1 at $\beta = 0$. A little reflection reveals that an intelligent modeling for $\alpha = A\gamma$ should have $(x_{21}, x_{22}, \ldots, x_{2k})'$ in $\underline{R}(A)$. Using the constant underlying hazard forces one into imbalanced modeling of the two treatment hazard functions, since treatment 1 will necessarily have constant hazard, while for any nonzero β treatment 2 must be nonconstant. This type of discrimination does not seem reasonable, for why should the shape of the two hazards be distinctly different? If one wishes to allow each hazard function the same possibilities within the modeling, then $(x_{21}, x_{22}, \ldots, x_{2k})'$ must be in $\underline{R}(A)$. So for balanced modeling the efficiency at $\beta = 0$ will be 1.

The results for this simple problem seem to indicate that the efficiency of the distribution-free approach with time dependent covariables will generally be very good against smooth, balanced parametric models. There are certainly many other practical designs of interest than just the ones considered in this work. However, the two-sample problem with

constant covariable, the two-sample problem with time dependent covariable, and the 2×2 factorial with nuisance β_2 seem to be basic types of designs, while most other designs can be regarded as extensions of these. The fact that the efficiency was very high in each of these three cases rather strongly suggests that for most practical designs of interest the distribution-free approach will be highly efficient against reasonable smooth parametric modeling.

IV. SUMMARY

The methods presented have been developed for analyzing factorial survival experiments in which the survival times are grouped by time periods. A grouped data version of the Cox regression model has been applied for the purpose of inference about treatment effects on survival time. The model is distribution-free in the sense that an underlying hazard function is left completely unrestricted; however, parametric regression expressions are used to explain treatment effects. These regression forms are very flexible, allowing the use of time dependent covariables to analyze effects which vary with time.

Large sample inference for the regression parameters β has been made practical by eliminating the nuisance parameters λ with an approximation. The true likelihood function is replaced by a good approximation, so that a maximum relative likelihood function $\ell^*(\beta)$, depending only on β, may be explicitly obtained. Large sample likelihood inference for β may then be carried out by treating $\ell^*(\beta)$ as a likelihood function for β. This allows one to quickly analyze various regressions models to find one that adequately fits and explains the data. After β has been estimated one can then easily estimate the nuisance parameters and the survival curves for each treatment. A toxicology experiment has been presented to illustrate the practicality of applying this distribution-free analysis.

Examination of the approximation reveals that it has been devised to give estimates very close to the true maximum likelihood estimates. The estimators obtained from the approximate likelihood are shown to be consistent, so they can be expected to work well for large sample

problems. There is a small gain in information caused by making the approximation. This will make estimates of $\text{Var}(\hat{\beta})$ and $\text{Var}(\hat{\lambda})$ slightly lower than they should really be; however, the hazard rates must be very high on the time periods for this approximation error to have an appreciable effect on the analysis. Proper experimental design to take more observations during time spans where many failures occur will alleviate this problem.

For most real data sets one would expect the hazard function to be relatively smooth. An important question arises in conjunction with the distribution-free underlying hazard function of the Cox model. Does leaving the underlying hazard unrestricted cause a loss in efficiency for inference about β compared to procedures based on smooth parametric modeling of the hazard functions? The grouped data setting provides a direct way to address this question by considering smoothing restrictions on the parameters α associated with the underlying hazard function. Exploitation of the connection between linear constraints $\alpha = A\gamma$ and the efficient score statistics makes it possible to express the asymptotic efficiency in terms of the known matrix A and the easily calculated information matrix obtained from the distribution-free model.

The two-sample problem was considered in depth, revealing that the asymptotic efficiency is 1 at $\beta = 0$ for any model which includes the constant hazard functions. Furthermore, the efficiency only slowly decreases as β moves away from 0, so that for interesting values of β the efficiency is always high. The geometric interpretation provides an intuitive way to view the efficiency problem and points out that the distribution-free analysis can be expected to do very well against

smooth models allowing the possibility of decreasing hazard functions. The distribution-free approaches for the 2×2 factorial problem and the two-sample problem with a time dependent covariable were also seen to be very efficient against smooth modeling. The results obtained indicate that the distribution-free approach is highly efficient for smooth underlying hazard functions, and the added scope of application for the distribution-free analysis seems well worth the minor loss in efficiency when smooth modeling is appropriate.

BIBLIOGRAPHY

Berkson, J., and R. P. Gage (1952). Survival curves for cancer patients following treatment. Journal of the American Statistical Association 47:501-515.

Bickel, P. J., and K. A. Doksum (1977). Mathematical Statistics: Basic Ideas and Selected Topics. San Francisco: Holden-Day.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. Journal of the Royal Statistical Society, Series B, 11:15-44.

Bradley, R. A., and J. J. Gart (1962). The asymptotic properties of ML estimators when sampling from associated populations. Biometrica 49:205-214.

Breslow, N. (1970). A generalized Kruskal-Wallis test for comparing samples subject to unequal patterns of censorship. Biometrika 57:579-594.

Breslow, N. (1972). Contribution to the discussion on the paper of D. R. Cox, Regression models and life-tables. Journal of the Royal Statistical Society, Series B, 34:216-217.

Breslow, N. (1974). Covariance analysis of censored survival data. Biometrics 30:89-99.

Breslow, N. (1975). Analysis of survival data under the proportional hazards model. International Statistical Review 43:45-57.

Chen, W., J. V. Fayos, and B. M. Hill (1977). Bayesian analysis of survival curves for cancer patients following treatments. Technical report, Department of Statistics, University of Michigan, Ann Arbor.

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society, Series B, 34:187-220.

Cox, D. R. (1975). Partial likelihood. Biometrika 62:269-276.

Cox, D. R., and D. V. Hinkley (1974). Theoretical Statistics. London: Chapman and Hall.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete samples via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39:1-22.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association, 72: 557-565.

Farewell, V. T., and R. L. Prentice (1977). A study of distributional shape in life testing. Technometrics, 19:69-75.

Feigl, P., and M. Zelen (1965). Estimation of exponential survival probabilities with concomitant information. Biometrics, 21:826-838.

Garton, R. R. (1975). Personal communication.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily single censored samples. Biometrika 52:203-223.

Glasser, M. (1967). Exponential survival with covariance. Journal of the American Statistical Association 62:561-568.

Gross, A. J., and V. A. Clark (1975). Survival Distributions: Reliability Applications in the Biomedical Sciences. New York: Wiley.

Holford, T. R. (1976). Life tables with concomitant information. Biometrics 32:587-597.

Kalbfleisch, J. D. (1974). Some efficiency calculations for survival distributions. Biometrika 61:31-37.

Kalbfleisch, J. D., and A. A. McIntosh (1977). Efficiency in survival distributions with time-dependent covariables. Biometrika 64: 47-50.

Kalbfleisch, J. D., and R. L. Prentice (1973). Marginal likelihoods based on Cox's regression and life model. Biometrika 60:267-278.

Kaplan, E. L., and P. Meier (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53:457-481.

Kendall, M. G., and A. Stuart (1973). The Advanced Theory of Statistics, Vol. 2. Third ed. London: Griffin.

Lee, E. T., M. M. Desu, and E. A. Gehan (1975). A Monte Carlo study of the power of some two-sample tests. Biometrika 62:425-432.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports 50:163-170.

Oakes, D. (1977). The asymptotic information in censored survival data. Biometrika 64:441-448.

Peto, R. (1972a). Rank tests of maximum power against Lehmann type alternatives. Biometrika 59:472-474.

Peto, R. (1972b). Contribution to the discussion on the paper of D. R. Cox, Regression models and life-tables. Journal of the Royal Statistical Society, Series B, 34:205-207.

Peto, R., and P. Lee (1973). Weibull distributions for continuous carcinogenesis experiments. Biometrics 29:457-570.

Peto, R., and J. Peto (1972). Asymptotically efficient rank invariant test procedures. Journal of the Royal Statistical Society, Series A, 135:185-206.

Prentice, R. L. (1973). Exponential survivals with censoring and explanatory variables. Biometrika 60:279-288.

Prentice, R. L., and L. A. Gloeckler (1978). Regression analysis of grouped survival data with application to breast cancer data. Biometrics 34:57-67.

Prentice, R. L., and E. R. Shillington (1975). Regression analysis of Weibull data and the analysis of clinical trials. Utilitas Mathematica 8:257-276.

Rao, C. R. (1973). Linear Statistical Inference and Its Applications. Second ed. New York: Wiley.

Richards, F. S. G. (1961). A method of maximum-likelihood estimation. Journal of the Royal Statistical Society, Series B, 23:469-475.

Thompson, W. A., Jr. (1977). On the treatment of grouped observations in life studies. Biometrics 33:463-470.

Zippin, C., and Armitage, P. (1966). Use of concomitant variables and incomplete survival information with estimation of an exponential survival parameter. Biometrics 22:665-672.