AN INVESTIGATION INTO THE OPERATING CHARACTERISTICS OF SOME TWO-SAMPLE NONPARAMETRIC TEST PROCEDURES USED FOR CENSORED SURVIVAL DATA

by

THOMAS R. FLEMING and DAVID P. HARRINGTON

Technical Report Series, No. 10
August 1980

An Investigation into the Operating Characteristics of Some Two-Sample Nonparametric Test Procedures Used for Censored Survival Data*

by

Thomas R. Fleming
Department of Medical Research Statistics of Epidemiology
Mayo Clinic
Rochester, Minnesota 55901

David P. Harrington
Department of Applied Mathematics and Computer Science
University of Virginia
Charlottesville, Virginia 22901

*This manuscript contains the results given in the contributed talk, "An Investigation of a Class of Kolmogorov-Smirnov-Type Test Procedures in Arbitiarily Right Censored Data," presented at the Joint Statistical Meetings of the American Statistical Association and the Biometric Society in Houston, Texas, on August 11, 1980.

This research was partially supported by the U.S. Department of Health Education and Welfare, Food and Drug Administration, through contract number 223-79-2274 awarded to EBON Research Systems.

Abstract

This report contains the results of investigations into the operating characteristics of various nonparametric test procedures used when examining censored survival data. The procedures are all two-sample test statistics, and include the Gehan-Wilcoxon, Log-Rank, and some new Smirnov-type statistics recently developed. These Smirnov-type statistics will be referred to as the Generalized Smirnov and $K^{\alpha}_{N_1,N_2}$ procedures. (N_1 and N_2 are the two sample sizes, and $\alpha \geq 0$ is a free parameter.)

Let S_1 and S_2 denote two survival distributions. When testing H_0: $S_1 = S_2$, theoretical considerations and Monte Carlo results support the conclusion that for $0 \leq \alpha < 1$, the $K^{\alpha}_{N_1,N_2}$ procedures have excellent sensitivity to detect crossing hazards departures from H_0 in which substantial survival differences exist later, but not earlier in time. Furthermore, $K^{\alpha}_{N_1,N_2}$ procedures for $\alpha \geq 2$ have excellent sensitivity to detect acceleration alternatives, that is, large early survival differences which disappear quickly in time. The Generalized Smirnov procedure turns out to be more versatile than the $K^{\alpha}_{N_1,N_2}$ procedures, providing good power generally against any of the crossing hazards alternatives examined. The Gehan-Wilcoxon and Log-Rank turn out to have relatively low power against most of the crossing hazards alternatives examined.

Table of Contents

0. Introduction

This report transmits all the results obtained by Thomas R. Fleming of the Mayo Clinic, Rochester, Minnesota and David P. Harrington of the University of Virginia, Charlottesville, Virginia on the project for Ebon Research Systems described in FDA Task Order Number 5. The primary purpose of the project was to evaluate the operating characteristics of some newly proposed test statistics useful in comparing two samples of censored survival data, and to compare these characteristics with those of certain statistics which have been in common use. The investigation was for the most part limited to underlying survival distributions with crossing hazard functions, i.e., survival distributions for which substantial differences evident at one point in time fail to exist at other points in time.

The outline of this report is as follows. Part I provides some general background information essential for understanding the specific numerical work done on this project. The new Smirnov-type test statistics we examined are defined in Part I, and the important known results about these statistics are summarized there. Part II describes the specific configurations of censoring and survival distributions that were used to produce Monte Carlo simulations of two-sample censored survival data; these simulations were used to evaluate the size and power of hypothesis tests based on the statistics studied. Part III contains a summary of the results of the simulations. Recommendations are given in Part III on how to pick the most sensitive test statistic, from among those considered, for detecting an anticipated difference

in two underlying survival distributions. Complete tables of all the simulation results can be found in Part IV. Parts V and VI contain references and an appendix, respectively.

I. Background Information

A detailed summary of the theoretical basis for much of the work done on this project can be found in the Preliminary Report submitted to Ebon. For the sake of brevity, we will only restate here the information from the Preliminary Report which is essential to understanding the results of the project.

A. Crossing Hazards Alternatives.

Suppose $X_{11}, X_{12}, \ldots, X_{1N_1}$ and $X_{21}, X_{22}, \ldots, X_{2N_2}$ are two independent samples of failure time random variables. These variables usually denote the time to a prespecified event (e.g., time to tumor progression) for each experimental unit in a study. In most survival studies, the failure time of each experimental unit may be censored, so let $(Y_{11}, Y_{12}, \ldots, Y_{1N_1})$ and $(Y_{21}, Y_{22}, \ldots, Y_{2N_2})$ denote the censoring times of the experimental units. For each experimental unit in the study, the observed data are usually $T_{ij} = \min (X_{ij}, Y_{ij})$ and $\delta_{ij} = I[X_{ij} \leq Y_{ij}]$, where $I[A] = 1$ if the event A occurs, and 0 otherwise; we will take this always to be the case. For simplicity, we will assume X_{ij} and Y_{ij} are statistically independent, although all results obtained continue to hold under the less stringent assumption detailed in Fleming and Harrington (1979).

Let $S_i(t) = P(X_{ij} > t)$, $i=1,2$; the most commonly encountered hypothesis test in the analysis of failure time data is H_0: $S_1(t) = S_2(t)$ for all t. If the alternative of interest is H_1: $S_1(t) < S_2(t)$ over some interval in t, the alternative is called one-sided. The general

alternative H_1: $S_1(t) \neq S_2(t)$ for some values of t is called two sided. The alternative hypotheses to the basic null hypothesis are clearly very complicated composite hypotheses. It is not realistic to expect that a single testing procedure would be adequately powerful against all alternatives of interest.

A particular type of alternative that may arise is called the "crossing hazards" alternative. If $\nu_i(t) = -\frac{d}{dt}\ell n\ S_i(t)$, i=1,2, then $\nu_i(t)$ is called the hazard rate or intensity function of the survival distribution $S_i(t)$. $\beta_i(t) = \int_0^t \nu_i(s)ds$ is called the cumulative hazard function, and it is well known that $S_i(t) = \exp[-\beta_i(t)]$. Now, when two underlying survival distributions have hazard functions which cross at some point, then the survival curves will exhibit differences over a time interval, but those differences may disappear outside that interval. For example, at a fixed value t_0 it is clearly possible that one might have $\beta_1(t_0) = \beta_2(t_0)$ (and hence $S_1(t_0) = S_2(t_0)$) even though $\beta_1(t) >> \beta_2(t)$ (and hence $S_1(t) << S_2(t)$) at some $t < t_0$. This will happen if the hazard functions cross at a point prior to time t_0 in such a way that the areas bounded by each of hazard functions and the time axis between $t = 0$ and $t = t_0$ are equal. This particular type of departure from the null hypothesis in which substantial early survival differences disappear later in time has been called the "acceleration alternative". The preliminary report for this project contains on page 2 a sketch of crossing hazard functions and the associated survival functions $S_i(t)$.

The crossing hazards phenomenon can often go undetected by test statistics that depend upon cumulative differences in the survival functions or, more specifically, cumulative differences in the hazard functions. The Gehan-Wilcoxon and the Log-Rank statistics are of this type. It is reasonable to expect, though, that procedures based upon maximum observed differences (perhaps weighted in some fashion) in empirical survival functions or empirical cumulative hazard rates might be more likely to detect crossing hazards alternatives to the null hypothesis H_0: $S_1(t) = S_2(t)$ for all t. Such procedures are usually called Kolmogorov-Smirnov-type (or just Smirnov-type) procedures because of the well known goodness-of-fit test based on the maximum observed difference between empirical and hypothesized cumulative distribution functions. Two kinds of Smirnov-type procedures have been proposed in the manuscripts by Fleming and Harrington (1979) and Fleming, O'Fallon, O'Brien, and Harrington (1979). (These manuscripts can be found in the appendixes of the Preliminary Report.) It is the sensitivity of these procedures that was investigated in this project. Specific definitions and properties of these test statistics are given in the next subsection.

B. The New Smirnov-Type Procedures.

The Preliminary Report gave a detailed account of these new Smirnov type procedures, including both a theoretical and heuristic discussion. We will limit ourselves here to careful definitions of the procedures, and a complete statement of the asymptotic distribution theory used to obtain significance levels of the test statistics.

The asymptotic distribution theory of the Smirnov-type statistics provides the most natural way to classify the statistics. The procedures described in both Fleming, et.al. (1979) and Fleming and Harrington (1979) are based on suprema of appropriately scaled empirical processes. The processes used in the first manuscript have asymptotic distributions which have the variance-covariance structure of a time transformation of a Brownian bridge, while those used in the second paper have asymptotic distributions of time transformations of a Brownian motion. We will discuss the Brownian bridge type procedure first.

1. Brownian Bridge Type Procedure.

Let X_{ij}, Y_{ij} and T_{ij} be the failure time, censoring time, and observed random variables, respectively, that were discussed earlier. The following notation was established in the Preliminary Report, but we review it here for the sake of completeness. Let:

$$S_i(t) = P(X_{ij} > t)$$

$$C_i(t) = P(Y_{ij} > t)$$

$$\pi_i(t) = P(T_{ij} > t)$$

$$\nu_i(t) = -\frac{d}{dt}\ell n\ S_i(t)$$

$$\gamma_i(t) = -\frac{d}{dt}\ell_n\ C_i(t)$$

$$\beta_i(t) = \int_0^t \nu_i(s)ds$$

$$\alpha_i(t) = \int_0^t \gamma_i(s)ds$$

$N_i(t)$ = number of experimental units in sample i still under observation just prior to time t (i.e., the size of the risk set in sample i at time t)

$D_i(t)$ = number of deaths observed in sample i at time t

$\delta_{ij} = I[X_{ij} \leq Y_{ij}]$ (I[A] is the usual indicator random variable of the event A.)

$\hat{\beta}_i(t) = \sum_{j:T_{ij} \leq t} [N_i(T_{ij})]^{-1} \delta_{ij}$. (This is the Nelson empirical cumulative hazard rate estimator of $\beta_i(t)$ for untied data.)

$\hat{\alpha}_i(t) = \sum_{j:T_{ij} \leq t} [N_i(T_{ij})]^{-1} (1-\delta_{ij})$. ($\hat{\alpha}_i(t)$ is the Nelson empirical estimator of $\alpha_i(t)$.)

$\hat{S}_i(t) = \exp[-\hat{\beta}_i(t)]$

$\hat{C}_i(t) = \exp[-\hat{\alpha}_i(t)]$

Observe that we have allowed the censoring distributions C_1 and C_2 to differ from one another.

We define the empirical process $Y_{N_1,N_2}(t)$ to be

$$Y_{N_1,N_2}(t) = \frac{1}{2}[\hat{S}_1(t) + \hat{S}_2(t)] \int_0^t \left[\frac{N_1\hat{C}_1(s-)\, N_2\hat{C}_2(s-)}{N_1\hat{C}_1(s-) + N_2\hat{C}_2(s-)} \right]^{1/2} d(\hat{\beta}_1(s) - \hat{\beta}_2(s)).$$

(Recall that N_1 and N_2 are the two sample sizes; we always take $f(s-) = \lim_{a \uparrow s} f(a)$ for any function f(s).)

The Preliminary Report discusses why we believe that a test statistic based on $\sup Y_{N_1,N_2}(t)$ should provide a particularly sensitive

test for detecting one- or two-sided crossing hazards type alternatives in situations where the underlying survival distributions exhibit their most substantial differences in the middle portion of the survival curves; i.e., at those values of t for which $S_i(t) \approx .5$. This conjecture is supported by the results summarized and tabulated in Sections III and IV. The calculation of approximate P-values using $\sup_t Y_{N_1,N_2}(t)$ is made possible by the following theorem, the proof of which may be found outlined in Fleming, et.al. In the statement of the theorem, "$\Rightarrow$" refers to weak convergence in $D[0,\tau]$, the space of functions on an interval $[0,\tau]$ with discontinuities of at most the first kind.

Theorem. Let $0 \leq t \leq \tau$, where τ is such that $\pi_i(\tau) > 0$, $i = 1,2$, and let $W = \{W(t): t \geq 0\}$ be a standard Wiener process. Let $S(t)$ be the common but unspecified value of $S_i(t)$, $i=1,2$, under H_0, and take $W_S(t)$ to be the time transformed Brownian bridge defined by

$$W_S(t) = W(1-S(t)) - [1-S(t)]W(1).$$

Then, under H_0,

$$\{Y_{N_1,N_2}(t): 0 \leq t \leq \tau\} \Rightarrow W_S \equiv \{W_S(t): 0 \leq t \leq \tau\}$$

as $N_1, N_2 \to \infty$ in such a way that $\lim_{N_1 \to \infty} N_1/N_2 = \lambda$, $0 < \lambda < \infty$.

The above weak convergence result implies that

$$\lim_{N_1,N_2 \to \infty} P\left[\sup_{0 \leq t \leq \tau} Y_{N_1,N_2}(t) > a\right] = P\left[\sup_{0 \leq t \leq \tau} W_S(t) > a\right],$$

The specific formula used to calculate the probability on the right hand side of the above equation, along with the computational algorithm used to calculate $\sup Y_{N_1,N_2}(t)$, can be found in Section 3.1.3 of the Preliminary Report.

2. Brownian Motion Type Procedure.

The notation established in the previous subsection holds here as well. In addition, we will need the following notation:

$$H_{N_1,N_2}(s) = \left[\frac{N_1\hat{C}_1(s-)\ N_2\hat{C}_2(s-)}{N_1\hat{C}_1(s-) + N_2\hat{C}_2(s-)}\right]^{\frac{1}{2}} \frac{1}{2}\{(\hat{S}_1(s-))^{\alpha} + (\hat{S}_2(s-))^{\alpha}\}$$

where α is a fixed nonnegative parameter. (Corresponding to each value of α will be a unique test procedure).

We define the empirical process $B^{\alpha}_{N_1,N_2}(t)$ to be

$$B^{\alpha}_{N_1,N_2}(t) = \int_0^t H_{N_1,N_2}(s)\, d(\hat{\beta}_1(s) - \hat{\beta}_2(s)),$$

and we let $B^{\alpha}_{N_1,N_2}$ denote the stochastic process $\{B^{\alpha}_{N_1,N_2}(t)\colon 0\le t\le\tau\}$.

The following asymptotic result is essential in formulating a Smirnov-type procedure based upon the process $B^{\alpha}_{N_1,N_2}$.

Theorem. Let $S(s)$ be the common value of $S_i(s)$, $i=1,2$, under H_0, and let $\{W(t),\ t\ge 0\}$ be a standard Brownian motion.

Then, under H_0,

$$B^{\alpha}_{N_1,N_2} \Rightarrow B^{\alpha} \equiv \{B^{\alpha}(t) = \int_0^t (S(s))^{\alpha-\frac{1}{2}}\,(\nu(s))^{\frac{1}{2}} dW(s)\colon\ 0\le t\le\tau\},$$

where τ is such that $\pi_i(\tau)>0$, $i=1,2$, and $N_1,N_2\to\infty$ so that $N_1/N_2\to\lambda$, $0<\lambda<\infty$.

If $(\hat{\sigma}_{\alpha}(t))^2$ is a consistent estimator of $\sigma^2_{\alpha}(t) \equiv \mathrm{Var}\, B^{\alpha}(t)$,

then the above result implies that $(\hat{\sigma}_{\alpha}(\tau))^{-1} B^{\alpha}_{N_1,N_2}(t)$, $0\le t\le\tau$, has,

for large sample sizes N_1 and N_2, approximately the distribution of a time transformed standard Brownian motion on [0,1]. Therefore, we have, for any value a, that

$$\lim_{N_1,N_2\to\infty} P\left[(\hat{\sigma}_\alpha(\tau))^{-1}\left[\sup_{0\le t\le\tau} B^\alpha_{N_1,N_2}(t) \ge a\right]\right] = P\left[\sup_{0\le u\le 1} W(u) \ge a\right] .$$

A Kolmogorov-Smirnov type procedure can therefore be based on the observed value of $K^\alpha_{N_1,N_2} \equiv (\hat{\sigma}_\alpha(\tau))^{-1} \sup_{0\le t\le\tau} B^\alpha_{N_1,N_2}(t)$, with significance levels computed according to the right hand side of the above equation. For reasons explained in the Preliminary Report, the particular consistent variance parameter estimate we have chosen is

$$(\hat{\sigma}_\alpha(\tau))^2 = \int_0^\tau [N_1\hat{C}_1(s-) + N_2\hat{C}_2(s-)]^{-1} \{\tfrac{1}{2}([\hat{S}_1(s-)]^\alpha + [\hat{S}_2(s-)]^\alpha)\}^2$$

$$\{N_2\hat{C}_2(s-)[\hat{S}_1(s-)]^{-1}d\hat{\beta}_1(s) + N_1\hat{C}_1(s-)[\hat{S}_2(s-)]^{-1}d\hat{\beta}_2(s).\}$$

The complexity of the statistic $K^\alpha_{N_1,N_2}$ appears at first glance a bit overwhelming. Each of its component pieces, however, can be easily motivated and such explanations can be found in pages 16-18 of the Preliminary Report. To understand the numerical results found in Sections III and IV it is essential only to be aware of the role α plays. α is a free parameter which is constrained to be nonnegative. If $\alpha>1$, $K^\alpha_{N_1,N_2}$ tends to emphasize nonzero values of the difference $\hat{S}_2(u) - \hat{S}_1(u)$ for those values u at which $S_i(u) \simeq 1$; such differences are often called early differences. The greater the value of α, the more emphasis placed on early differences. Such an emphasis, however, will always cause a corresponding de-emphasis of differences observed at

other time points, and the larger the value of α, the more $K^{\alpha}_{N_1,N_2}$ will discount differences in $\hat{S}_2(u) - \hat{S}_1(u)$ at points where $S_i(u) << 1$, $i=1,2$. Procedures based on small values of $\alpha<1$, on the other hand, emphasize changes in the difference $\hat{S}_2(u) - \hat{S}_1(u)$ which occur when $S_i(u) \simeq 0$, $i=1,2$, i.e., differences which are said to occur later in time.

The qualitative role of α is supported by both the asymptotic theory and heuristic explanations of the test statistic (see Preliminary Report). Until this project, however, we had very little intuition about how large or small α must be to provide acceptable power against specific instances of crossing hazards alternatives. Although we are still a long way from a complete quantitative understanding of the role of α, the results tabulated in the next two sections provide a very good beginning at establishing guidelines for a judicious choice of α.

We feel it is important to emphasize a point here regarding the choice of α. The parameter α is a component of the statistic $K^{\alpha}_{N_1,N_2}$ that should be specified by a researcher in advance of seeing the data. If a data analyst chooses to use $K^{\alpha}_{N_1,N_2}$ and feels that it is of utmost importance to detect differences in underlying survival distributions which occur early in time, then α should be chosen as large as is prudent ($\alpha = 2$ is nearly always large enough.) To examine the data first, however, before choosing α would be irresponsible "data dredging", since it is clear that with a clever choice of α, very many data sets can be shown to contain statistically significant differences between underlying survival distributions.

Both the Brownian motion and the Brownian bridge based procedures are clearly complex statistics. The asymptotic distribution theory only tells us how to construct hypothesis tests of a given size; analytic power calculations seem nearly impossible at this stage. Monte Carlo simulations seem to be the only manageable means of determining the power of these procedures in some representative situations. Furthermore, the simulations provide a method to determine if the true size of these test procedures in small and moderate samples is accurately approximated by the nominal significance level based upon the appropriate asymptotic distribution theory. The configurations of censoring and survival distributions used to produce the simulations are briefly described in the next section, and specified in detail in Section IV. All random variables generated in the configurations were produced by transforming uniform random variables generated with the linear congruential method (Knuth, 1969).

The exact formulas for the hazard rates of the survival distributions and for the censoring distribution functions employed in generating the censored survival data are given in Section IV with the tabulated results. We feel it is important, however, to explain the general strategy used in choosing the specific distributions, and to give a summary of the kinds of distributions chosen. The reader will then be able to judge Section III, The Qualitative Summary of the Results, more critically.

We used seventeen distinct configurations of survival and censoring distributions in all, with each configuration including two survival distributions used to generate the two independent samples of failure times, and a single censoring distribution used to generate the two independent samples of censoring times. All censoring and survival random variables were generated independently, with each observation time taken to be the minimum of a survival and a censoring random variable; that is, $T_{ij} = \min (X_{ij}, Y_{ij})$ (as indicated earlier). The sample sizes N_1 and N_2 of the two independent samples used for testing H_0: $S_1 = S_2$ were taken to be equal for a given simulation. For each configuration two distinct values of the common sample size N_i were inspected. Five hundred pairs of samples (one thousand pairs of samples when evaluating size) were generated for each selected configuration of survival and censoring distributions for the two populations and for each sample size. The proportions of samples in which each one-sided test procedure under consideration rejected H_0 at the $\alpha = 0.01$ and

$\alpha = 0.05$ significance levels were calculated for each configuration at each sample size.

In all except two cases, the survival distributions chosen possessed piecewise constant hazard rates, and thus were piecewise exponential distributions. The two exceptions were configurations 8 and 12 which contained one or more Weibull survival distributions with a shape parameter different from one. Semi-logarithmic plots of the survival functions can be found in Section IV with the tabled results.

The configurations chosen fell into three main categories:

1. The null hypothesis class of distributions, i.e., configurations in which $S_1 = S_2$.

2. Representative classes of either commonly arising crossing hazards alternatives, or proportional hazards alternatives.

3. Distributions which could reasonably be considered to have generated the FDA 165-174 or 165-150 mouse study data. These configurations enabled us to evaluate the power of the Smirnov-type procedures as well as the power of the Gehan-Wilcoxon and the Log-Rank procedures in situations that were of particular interest to the FDA. We will now summarize the kinds of configurations used in each of the above categories.

Of the seventeen configurations used, the first six fell into category 1. In configurations 1, 3 and 5, equal exponential survival distributions with constant hazard rates $\lambda = 2, 1$ and 0.5 respectively were used with a censoring distribution that produced only terminal censoring, that is, $Y_{ij} = \tau$, a constant, for all i and j. Configurations

2, 4 and 6 were generated using the same three exponential survival distributions listed above. Here, however, the censoring distribution was chosen to be a truncated uniform distribution (see Figure 4.2) which was selected to replicate as closely as possible the type of censoring distribution that was observed in the time-to-RE-tumor data of FDA study 165-174. With this approach, we were able to inspect the true size of the various test procedures in data which was lightly, moderately or heavily censored; specifically, the expected percents censored in configurations 1 through 6 were 13%, 25%, 37%, 47%, 61% and 68% respectively. In category 1, configuration 6 most nearly approximates the actual configuration seen in FDA study 165-174, and hence enables us to inspect the true sizes of the procedures in the actual setting in which we are currently most interested. In each of the first six configurations, simulations were performed separately, first for $N_1 = N_2 = 20$, and then for $N_1 = N_2 = 50$, since the intent was to inspect in small and moderate sample sizes the behavior of procedures whose significance levels were determined using appropriate asymptotic results.

Configurations 7 through 12 fell into category 2. Each of these configurations had the truncated uniform censoring distribution identical to that employed in configurations 2, 4 and 6. Configuration 7 presents a "proportional hazards" or "Lehmann" alternative. Specifically, two exponential distributions representing a doubling in median survival were generated. This configuration was chosen to enable us to compare the behavior of the Smirnov-type procedures to that of the Log-Rank in the situation in which the latter test procedure would be expected

to have its greatest relative sensitivity. (see Peto & Peto (1972)).

Configurations 8, 9 and 12 present departures from the null hypothesis in which substantial differences existing between survival distributions later in time fail to exist early in time. By inspecting the formulas for the Gehan-Wilcoxon and Log-Rank test statistics, as we will do in Part III, it is quite clear that the Gehan-Wilcoxon procedure will have unacceptable power and the Log-Rank procedure generally marginally acceptable power to detect this type of crossing hazards alternative. Configuration 8 used two Weibull distributions in which $S_2(t) >> S_1(t)$ for large t even though $S_2(t)$ is slightly less than $S_1(t)$ for $t \simeq 0$. This type of departure from H_0 could be expected to arise when one is comparing the survival of aggressively treated patients with coronary heart disease to that of patients treated more conservatively. Configuration 9, comprised of two piecewise exponential distributions, is very similar in form to configuration 8 except for the fact that $S_1(t) = S_2(t)$ for small t. Thus, configuration 9 will enable us to determine whether any additional power the Smirnov-type procedures may have over the Log-Rank procedure in configuration 8 will still exist in a situation in which $S_1(t) \leq S_2(t)$ for all t and in which the hazard functions technically don't cross. Configuration 12 is again similar to configuration 8. However it uses two Weibull distributions, one with an increasing and one with a decreasing hazard function, having enormous survival differences later in time.

Configurations 10 and 11 both present crossing hazards alternatives to the null hypothesis where all survival distributions are piecewise exponential. In configuration 10, large differences exist between survival curves over the middle range of the survival distribution although $S_1 = S_2$ for both small t and large t. Configuration 11 presents the situation in which large early differences between survival curves disappear somewhat later in time. These types of departures from the null hypothesis, sometimes referred to as "acceleration alternatives", are commonly observed when one is comparing survival or time to progression of disease curves for two chemotherapeutic or radiation therapy anti-tumor regimens in perspectively randonized clinical trials. From the formulation of their test statistics, we would anticipate the Log-Rank procedure to have unacceptable sensitivity to these departures, while the Gehan-Wilcoxon procedure should have marginally acceptable power against configuration 11. Here, as throughout configurations 1 through 12, we inspected both small and moderate sample size behavior, that is, we generated sample sizes $N_1 = N_2 = 20$, and then $N_1 = N_2 = 50$.

The last five configurations (13 through 17) are members of category 3. The data from mouse study 165-174 was used to construct survival and censoring distributions in 13, 14 and 15. The time scale was taken so that 1 unit = 100 weeks. The censoring pattern was essentially the same as the one used in configurations 2, 4, 6 and 7 through 12. Specifically, the censorship distribution was a truncated uniform distribution having a lag time of 60 weeks and complete censorship

at 111 weeks (see Figure 4.15). This distribution was chosen since it was found to very nearly approximate the Kaplan-Meier estimates of the censoring distributions for both the female control group 1 and the female high dose Red dye #40 group in the time-to-RE-tumor data for study 165-174. Configuration 13 used piecewise exponential survival models to approximate the actual departure from the null hypothesis that was observed in the female mice from study 165-174 when Kaplan-Meier estimates of time-to-RE-tumor curves were generated for the pooled control groups and then for the low dose Red dye #40 group (see Figure 4.16). The maximum difference of 0.12 between these curves occurs at t = 1.08. Configuration 14 used similar piecewise exponential survival models, but enlarged the maximum difference at t = 1.08 to 0.20. In configuration 15, this difference was enlarged still further to a difference of 0.27. The survival curves in 15 were each within reasonable confidence bands which could be constructed about the corresponding Kaplan-Meier estimated time-to-RE-tumor curves given in Figure 4.16. The sequence of configurations 13 through 15 allows us to examine the dependence of the power functions of Smirnov-type, Gehan-Wilcoxon and Log-Rank procedures on the degree of difference in survival distributions for crossing hazards alternatives of this type.

Configurations 16 and 17 were modeled after the 165-150 mouse study. The censorship distribution was a truncated uniform distribution which would have closely approximated the actual censoring distributions in the control, low dose and medium dose groups of female mice if no

interim sacrifice had been performed (see Figure 4.17). Configuration 16 used piecewise exponential survival curves to approximate the actual departure from the null hypothesis that was observed in the female mice from study 165-150 when Kaplan-Meier estimates of time-to-RE-tumor curves were generated for the control group and then for the pooled low, medium and high dose Red dye #40 groups (see Figure 4.18). The maximum difference of 0.11 between the curves occurs at $t = 0.91$. Configuration 17 enlarged the observed maximum difference of 0.11 to 0.17 at $t = 0.91$ to examine, as before, the change in power caused by a change in the true difference between the survival curves.

In mouse study 165-150, 50 animals of each sex were entered in each dosage group. Twice that number were entered in study 165-174. Hence in configurations 13 through 17, simulations were performed separately, first for $N_1 = N_2 = 50$, and then for $N_1 = N_2 = 100$. This, for example, enables power calculations for the situations in study 165-150 in which pooling by sex was not and was done respectively.

It should be noted that the intent in configurations 13 through 17 of our Monte Carlo investigation was not to prove or disprove that substantial evidence exists to support a hypothesis concerning the carcinogenicity of Red dye #40. Rather, our intent was solely to evaluate for future experiments the general behavior of certain test procedures. Specifically, we wanted to compare their ability to detect certain meaningful types of crossing hazards alternatives to the null hypothesis that may have truly existed in the Red dye #40 mouse experiments.

III. Qualitative Summary of the Results

A. Gehan-Wilcoxon and Log-Rank Test Statistics

Before discussing the results of the Monte Carlo simulations, it will be useful to briefly review the general form of the Gehan-Wilcoxon and Log-Rank two sample test statistics. For simplicity we will momentarily assume no ties exist in the data. Previous authors, including Prentice and Marek (1979), have observed that the Log-Rank test statistic can be formulated as

$$(\hat{\sigma}^2_{LR})^{-\frac{1}{2}} \sum_{j=1}^{d} \{D_1(T_j) - \frac{N_1(T_j)}{N_1(T_j) + N_2(T_j)}\} \tag{4.1}$$

where $\{T_j: j=1, \ldots,d\}$ is the set of d distinct observed death times in the pooled sample, and $\hat{\sigma}_{LR}$ is an appropriate variance estimator. Furthermore, the Gehan-Wilcoxon test statistic can be formulated as

$$(\hat{\sigma}^2_{GW})^{-\frac{1}{2}} \sum_{j=1}^{d} \{N_1(T_j) + N_2(T_j)\} \{D_1(T_j) - \frac{N_1(T_j)}{N_1(T_j) + N_2(T_j)}\} \tag{4.2}$$

where again $\hat{\sigma}^2_{GW}$ is an appropriate variance estimator.

Inspection of (4.1) reveals that the Log-Rank test statistic can essentially be viewed as a weighted difference, where the difference is between the "total observed deaths" in one sample and that sample's "total expected deaths given H_0 holds". Now, on the average the observed number of deaths in sample i will exceed the expected number of deaths under H_0 in any interval in which population i has the greater hazard function. The reverse will hold over intervals in which population i has the smaller hazard. Therefore, one would not anticipate that the Log-Rank test will be particularly

sensitive to crossing hazards alternatives. For similar reasons, inspection of (4.2) leads one to speculate that the Gehan-Wilcoxon test procedure also will lack sensitivity to that type of departure from H_0.

Interestingly, because the Gehan-Wilcoxon statistic differs from the Log-Rank statistic primarily because of its weighting factor $\{N_1(T_j) + N_2(T_j)\}$ (see (4.2)), we anticipate the Gehan-Wilcoxon procedure will have greater sensitivity than the Log-Rank procedure to departures from H_0 which are most evident early in time. However, the Log-Rank will have the greater sensitivity to those differences most evident later in time.

B. Results of Category 1 Simulations: Size

Results of simulations for all configurations 1 through 17 appear in Tables 4.1 through 4.17 respectively. In each configuration the behavior of eight one-sided test procedures were inspected; specifically, the Smirnov-type procedure based upon an underlying Brownian bridge process (hereafter exclusively referred to as the Generalized Smirnov procedure), the Smirnov-type procedures based upon an underlying Brownian motion process and corresponding to $\alpha = 0,1,2,3$ and 4 (procedures hereafter referred to as $K^{\alpha}_{N_1,N_2}$ procedures), and finally the Gehan-Wilcoxon and Log-Rank procedures.

Results pertaining to size of these procedures are presented in Tables 4.1 through 4.6 respectively. Overall, the Generalized Smirnov procedure comes very close to the nominal 0.01 level at both $N_i = 20$ and $N_i = 50$, but is slightly conservative at the 0.05 nominal level. In comparison the $K^{\alpha}_{N_1,N_2}$ procedures for $\alpha = 1,2,3$ and 4

are quite conservative at $N_i = 20$, but comparable in size to the Generalized Smirnov procedure in samples of size $N_i = 50$. Interestingly, the $K^0_{N_1,N_2}$ procedure is very conservative at both $N_i = 20$ and $N_i = 50$, much like the small sample behavior of classical Kolmogorov-Smirnov statistics in uncensored data.

C. Results of Category 2 Simulations:

General Crossing Hazards or Proportional Hazards Alternatives.

In Table 4.7 it is clear that the Log-Rank test procedure is, as we would anticipate, most sensitive in detecting proportional hazard alternatives. However, its gain in power is not large. For example, when $N_i = 50$ and the nominal level is 0.05, the power of the Log-Rank is 0.87, of the $K^1_{N_1,N_2}$ is 0.83, of the Gehan-Wilcoxon is 0.82 and of the Generalized Smirnov is 0.80. Interestingly, of all the $K^\alpha_{N_1,N_2}$ procedures considered, $K^1_{N_1,N_2}$ is the most powerful against the Lehmann alternative.

Tables 4.8, 4.9 and 4.12 present results for departures from the null hypothesis in which substantial differences existing between survival distributions later in time fail to exist early in time. As we anticipated, the Log-Rank has marginally acceptable power against these alternatives, far better than the unacceptable power of the Gehan-Wilcoxon procedure. In turn, however, the Generalized Smirnov procedure has power clearly better than that of the Log-Rank. The power of the $K^\alpha_{N_1,N_2}$ procedures to detect these later differences depends dramatically upon the choice of α. The procedure based upon $K^0_{N_1,N_2}$ is the most sensitive

of all eight test procedures in each of the three configurations. Possibly the most interesting of the three configurations is #9 since here $S_2(t) \geq S_1(t)$ for all t. When $N_i = 50$ and looking at the 0.05 level, the power of the Gehan-Wilcoxon is only 0.25, compared to 0.69 for the Log-Rank and 0.85 for the Generalized Smirnov, a marked reversal of the relative power of the latter two procedures from that which existed in the proportional hazards setting. The $K^0_{N_1,N_2}$, $K^2_{N_1,N_2}$ and $K^4_{N_1,N_2}$ procedures had powers of 0.95, 0.35, and 0.10 respectively, providing clear evidence of the powerful effect of the free parameter α.

The Generalized Smirnov procedure is unquestionably the most sensitive procedure in detecting large differences between survival curves over the middle range of the survival distribution, as shown in Table 4.10. When an "acceleration alternative" exists, that is, when large early differences between survival distributions disappear later in time as in Figure 4.11, the Generalized Smirnov procedure again is considerably more sensitive than both the Gehan-Wilcoxon and Log-Rank procedures. The powers of these three procedures for $N_i = 50$ at the 0.05 level are 0.82, 0.52, and 0.21 respectively. Further, $K^0_{N_1,N_2}$ and $K^2_{N_1,N_2}$ had powers of 0.12 and 0.84 respectively, again providing clear evidence of the dramatic effect the parameter α can have in the ability of $K^\alpha_{N_1,N_2}$ to detect crossing hazards departures from H_0. The power of $K^\alpha_{N_1,N_2}$ to detect "acceleration alternatives" is substantially increased by choosing larger α values, as we concluded

earlier from theoretical considerations.

D. Results of Category 3 Simulations:

Acceleration Alternatives in FDA Mouse Studies

Results of configurations 13 through 15, modeled after data from mouse study 165-174, and results of configurations 16 and 17, modeled after data from mouse study 165-150, are presented in Tables 4.13 through 4.17.

These results clearly confirm earlier conclusions, based upon theoretical considerations, that the Log-Rank procedure has unacceptable sensitivity and the Gehan-Wilcoxon only marginally adequate sensitivity to detect "acceleration alternatives". In fact the power of the Gehan-Wilcoxon becomes considerably less acceptable relative to the power of the Generalized Smirnov or $K^{\alpha}_{N_1,N_2}$ (for α = 2,3 or 4) procedures as the magnitude of the acceleration alternative increases.

When dealing with samples of size 100 (as would be the case in study 165-150 with pooling by sex and in study 165-174 if pooling by sex is not performed) and using α = 0.05 level tests, both the Generalized Smirnov and $K^{\alpha}_{N_1,N_2}$ (for α = 2,3 and 4) procedures appear to have reasonably good power to detect the type of acceleration alternatives seen in studies 165-174 and 165-150 if the maximum separation between curves is at least 0.17 to 0.20. The results of these specific configurations are given in Table 4.14 for study 165-174 and in Table 4.17 for study 165-150.

Data presented in Tables 4.13 through 4.17 confirm earlier theoretical conclusions that $K^{\alpha}_{N_1,N_2}$ procedures with $\alpha > 1$ have much

greater power to detect large early survival differences which disappear later in time than procedures with $\alpha \leq 1$. It is of interest to look more closely at results in Tables 4.14 and 4.17. As stated earlier, they present two acceleration alternatives in which the maximum separation between curves is of the same order of magnitude. However, in Table 4.14 the increase in power corresponding to an increase in α is less than that observed in Tables 4.17 for an equivalent increase in α. This is due to the fact that the maximal separation between curves occurs "sooner" in configuration 17 than in configuration 14, specifically 0.97 vs 0.80 as opposed to 0.83 vs 0.63. This provides further testimony to the dramatic ability of the parameter α to determine precisely over what intervals the procedure $K^{\alpha}_{N_1,N_2}$ has its greatest sensitivity to detect departures from H_0.

E. General Recommendations

From the results obtained from theoretical considerations as well as Monte Carlo simulations, it is quite clear that $K^{\alpha}_{N_1,N_2}$ procedures for $\alpha < 1$ have excellent sensitivity, unsurpassed by any other two-sample test procedures considered, to detect crossing hazards departures from H_0 in which substantial survival differences exist later, but not earlier, in time, (see, for example, Tables 4.8, 4.9 and 4.12). Furthermore, $K^{\alpha}_{N_1,N_2}$ procedures for α of two or greater have excellent sensitivity to detect acceleration alternatives, that is, large early survival differences which disappear quickly in time, (see, for example, Tables 4.11 and 4.13 through 4.17).

Unfortunately, as one might expect, individual $K^{\alpha}_{N_1,N_2}$ procedures for $\alpha \neq 1$ lack the versatility of having good power against crossing hazards alternatives of all forms. Specifically $K^{0}_{N_1,N_2}$ has relatively low power in configurations 11 and 13 through 17, while $K^{\alpha}_{N_1,N_2}$ for $\alpha = 2,3$ or 4 has low power in configurations 8, 9 and 12. However, this versatility of good power against any substantial crossing hazards departure from H_0 certainly is a property of the Generalized Smirnov procedure. In every configuration 8 through 17, the Generalized Smirnov procedure has either the best or close to the best power of any of the eight procedures considered. For this reason, we would generally recommend that the Generalized Smirnov procedure be employed when one is interested in detecting crossing hazards alternatives to H_0, including the "acceleration alternative".

We hasten to point out that we do not mean to imply that the Generalized Smirnov procedure is always superior to the Gehan-Wilcoxon or Log-Rank procedures relative to any departure from the null hypothesis. The latter two procedures are classical procedures each having been shown to be very powerful in their abilities to detect certain types of differences between survival distributions. Hence, the Generalized Smirnov procedure should be viewed as complementing the Gehan-Wilcoxon and Log-Rank procedures and not as a competitor. The procedure one chooses to use to test for the equality of two survival distributions will therefore depend upon the type of distributional differences for which one desires particular sensitivity. In conclusion the

Generalized Smirnov procedure would appear to be the appropriate choice when one wishes to have sensitivity to differences which are large at some point in time, independent of the type of differences existing elsewhere. Thus, it would appear from our results that the Generalized Smirnov procedure would be the most appropriate of the procedures which we have considered to test for substantial "acceleration alternatives" to the null hypothesis.

IV. Tabled Results of the Simulations

In tabulating the results of the simulations, we have opted for clarity at the expense of economy. The following pages contain seventeen tables, numbered 4.1 - 4.17; each table presents the simulation results (estimated size or estimated power) for a single configuration of censoring and survival distributions. Each tabled value is the observed proportion of times that the indicated test statistic produced a significance level less than or equal to the given nominal significance level (either $\alpha = .01$ or $\alpha = .05$). Appropriate sample sizes are indicated in the table headings, and the number of replications or simulations used in computing the proportion of rejections of H_0 is given at the top of the page. The Brownian bridge type procedure is referred to as the Generalized Smirnov procedure, while the specific choices of the Brownian motion procedure are labeled K^{α}, $\alpha = 0,1,2,3$ and 4.

Tables 4.1 through 4.12 each appear on a separate page, with graphs of the relevant censoring and survival distributions given in a figure just above each table. The graphs for the simulations based on the FDA mouse study data are a bit more complex, and were thus displayed separately. Figure 4.13 shows the three separate configurations of survival and censoring distributions from the 165-174 mouse study data. So as not to clutter the graphs, important values of the hazard and survival functions are given on the next page, while the three pertinent power tables (4.13 - 4.15) follow on the next two pages. The distributions estimated from the 165-150 mouse study data are shown in Figure 4.14; power Tables 4.16 and 4.17 again follow the page of hazard function and survival function values.

The last four graphs of this report follow Tables 4.16 and 4.17. These graphs are labeled Figures 4.15 through 4.18 and they show the Kaplan-Meier estimates of censoring and survival curves from the FDA mouse study data used to construct the distributions for configurations 13 through 17. Figures 4.15 and 4.16 show the empirical censoring and survival curves, respectively, referred to earlier for project number 165-174. The censoring distribution used in configurations 13 through 15 is superimposed on Figure 4.15, while the survival distributions used in configuration 13 are shown on Figure 4.16. Figure 4.17 shows both the Kaplan-Meier estimate of the censoring pattern for the relevant data in the 165-150 study, and the censoring distribution we chose for configurations 16 and 17. Figure 4.18 displays empirical survival curves for part of the 165-150 data, and the piecewise exponential survival distributions used in simulation configuration 16.

Monte-Carlo Estimates of the Sizes of the Generalized Smirnov, K^{α} (α = 0,1,2,3,4), Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of H_0: $S_1 = S_2$ vs H_1: $S_1 < S_2$ (1000 simulations)

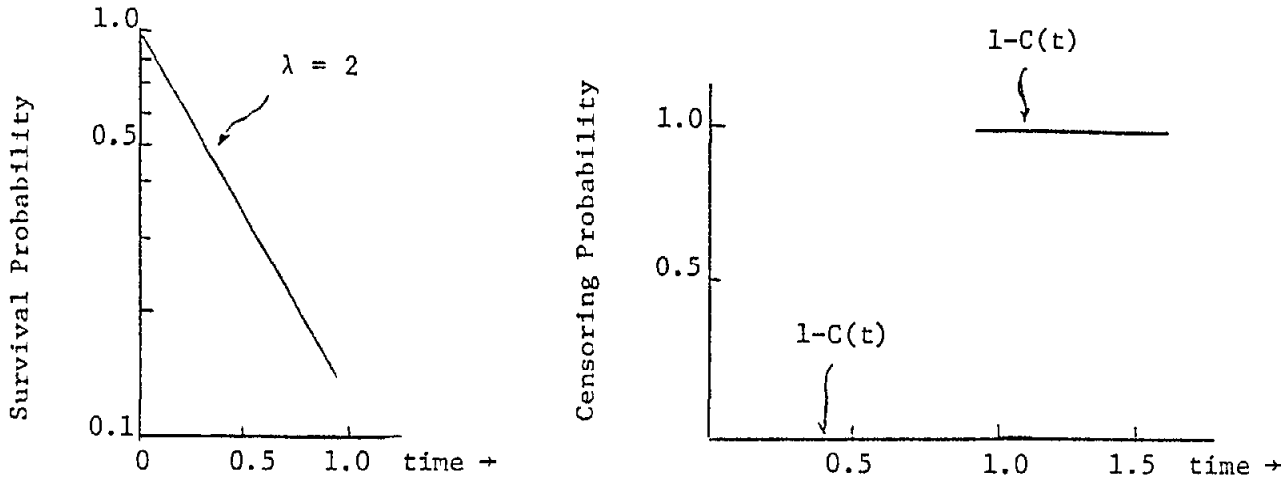


FIGURE 4.1: CONFIGURATION 1

Expected Percent Censored: 13.5%

Sample Size:	$N_1 = N_2 = 20$		$N_1 = N_2 = 50$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.012	.044	.010	.039
K^0	.001	.029	.002	.028
K^1	.003	.029	.007	.039
K^2	.004	.035	.013	.042
K^3	.006	.038	.011	.046
K^4	.006	.047	.012	.051
Gehan-Wilcoxon	.008	.042	.008	.047
Log-Rank	.006	.043	.011	.046

TABLE 4.2 Size

Monte-Carlo Estimates of the Sizes of the Generalized Smirnov, K^{α} (α = 0,1,2,3,4), Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of H_0: $S_1 = S_2$ vs H_1: $S_1 < S_2$ (1000 simulations)

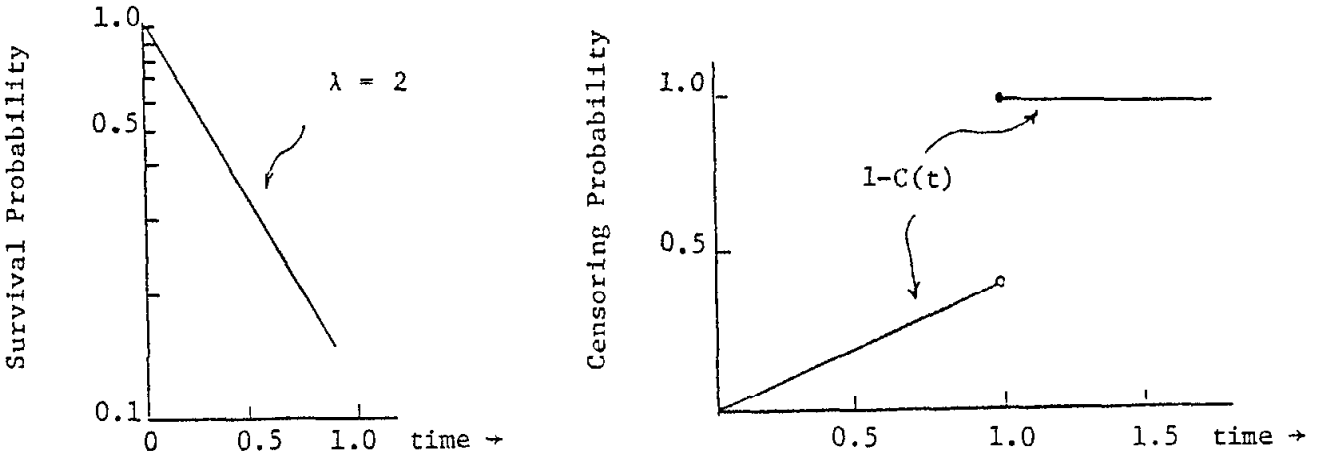


FIGURE 4.2: CONFIGURATION 2

Expected Percent Censored: 25.4%

Sample Size:	$N_1 = N_2 = 20$		$N_1 = N_2 = 50$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.014	.057	.012	.051
K^0	.000	.020	.003	.036
K^1	.003	.033	.009	.048
K^2	.004	.037	.006	.051
K^3	.004	.042	.005	.047
K^4	.003	.042	.002	.040
Gehan-Wilcoxon	.007	.056	.011	.054
Log-Rank	.007	.045	.011	.050

TABLE 4.3 SIZE

Monte-Carlo Estimates of the Sizes of the Generalized Smirnov, K^{α} (α = 0,1,2,3,4), Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of H_0: $S_1 = S_2$ vs H_1: $S_1 < S_2$ (1000 simulations)

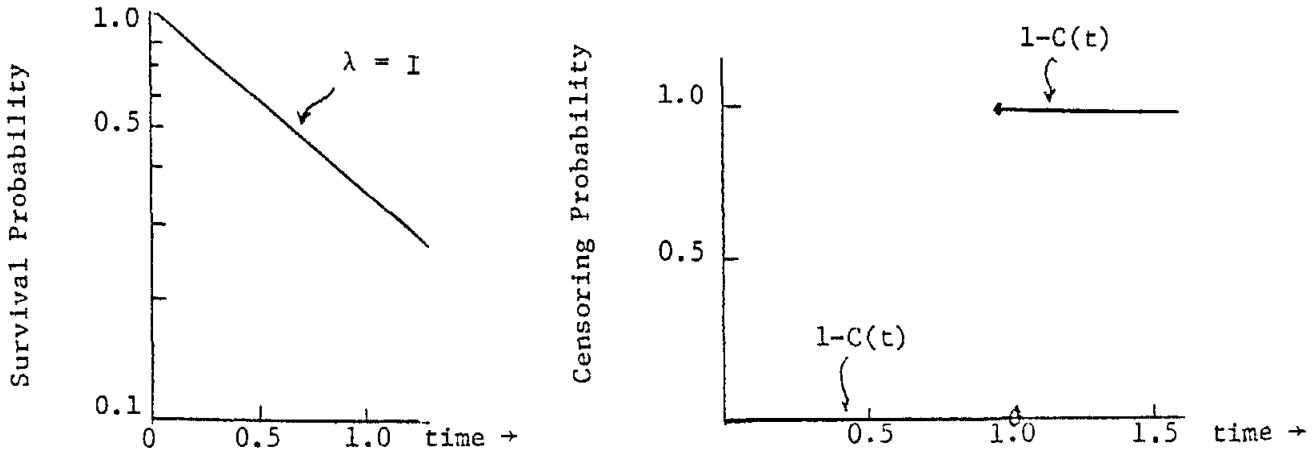


FIGURE 4.3: CONFIGURATION 3

Expected Percent Censored: 36.8%

Sample Size:	$N_1 = N_2 = 20$		$N_1 = N_2 = 50$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.006	.036	.009	.040
K^0	.000	.033	.003	.035
K^1	.000	.033	.004	.035
K^2	.000	.033	.004	.034
K^3	.002	.033	.003	.035
K^4	.002	.038	.004	.033
Gehan-Wilcoxon	.003	.058	.004	.033
Log-Rank	.006	.060	.006	.040

TABLE 4.4 SIZE

Monte-Carlo Estimates of the Sizes of the Generalized Smirnov, K^{α} (α = 0,1,2,3,4), Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of H_0: $S_1 = S_2$ vs H_1: $S_1 < S_2$ (1000 simulations)

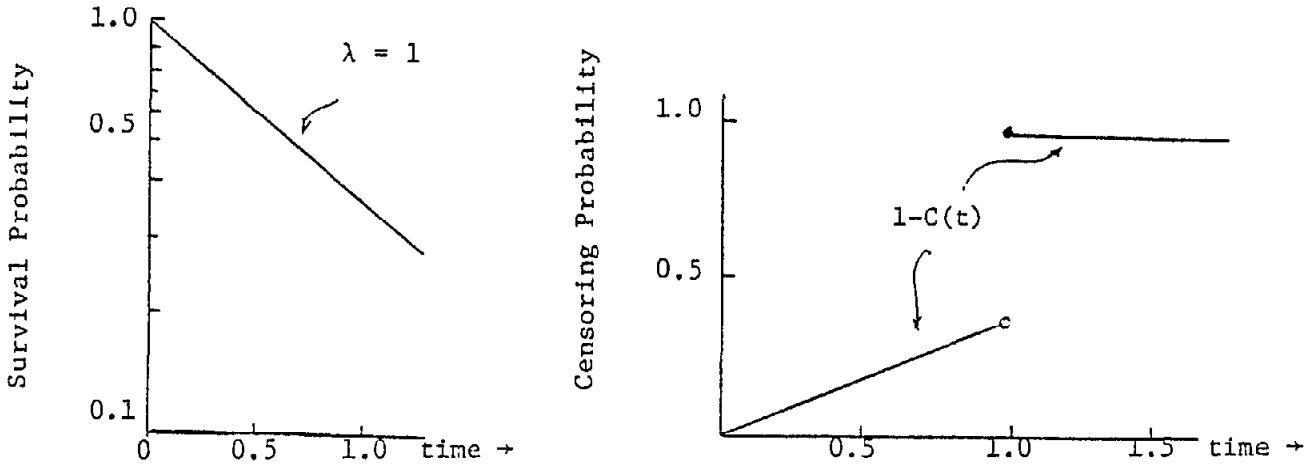


FIGURE 4.4: CONFIGURATION 4

Expected Percent Censored: 47.4%

Sample Size:	$N_1 = N_2 = 20$		$N_1 = N_2 = 50$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.007	.038	.008	.041
K^0	.001	.023	.007	.035
K^1	.002	.034	.010	.043
K^2	.003	.035	.008	.045
K^3	.005	.030	.006	.045
K^4	.005	.033	.006	.036
Gehan-Wilcoxon	.004	.048	.011	.050
Log-Rank	.004	.049	.015	.054

Monte-Carlo Estimates of the Sizes of the Generalized Smirnov, K^{α} (α = 0,1,2,3,4), Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of H_0: $S_1 = S_2$ vs H_1: $S_1 < S_2$ (1000 simulations)

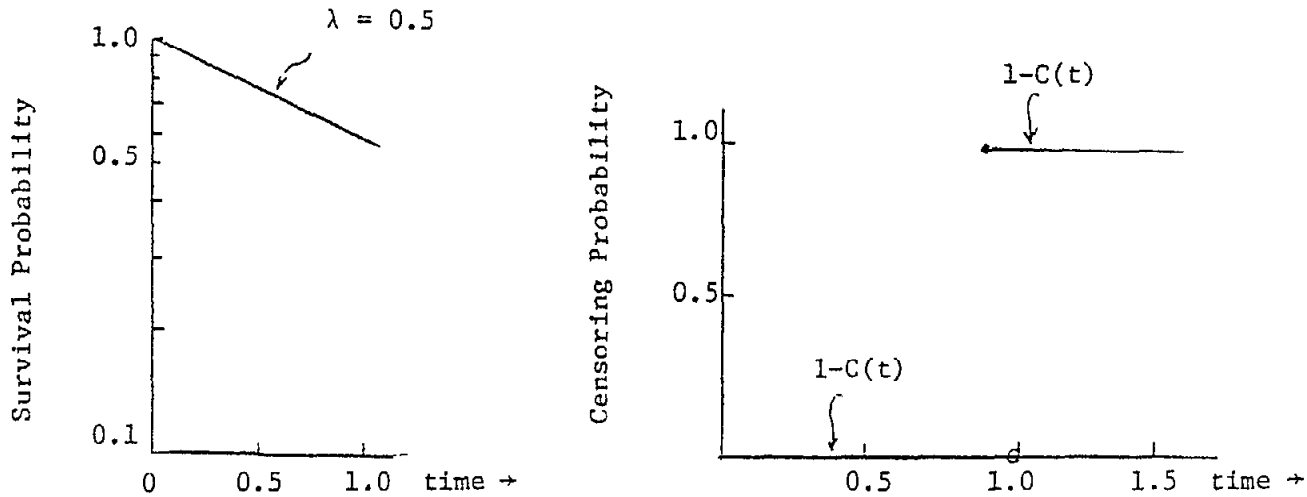


FIGURE 4.5: CONFIGURATION 5

Expected Percent Censored: 60.7%

Sample Size:	$N_1 = N_2 = 20$		$N_1 = N_2 = 50$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.011	.045	.007	.039
K^0	.004	.036	.005	.042
K^1	.004	.042	.006	.040
K^2	.005	.043	.006	.036
K^3	.005	.044	.006	.038
K^4	.005	.042	.005	.038
Gehan-Wilcoxon	.010	.061	.012	.046
Log-Rank	.009	.064	.011	.049

TABLE 4.6 SIZE

Monte-Carlo Estimates of the Sizes of the Generalized Smirnov, K^{α} (α = 0,1,2,3,4), Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of H_0: $S_1 = S_2$ vs H_1: $S_1 < S_2$ (1000 simulations)

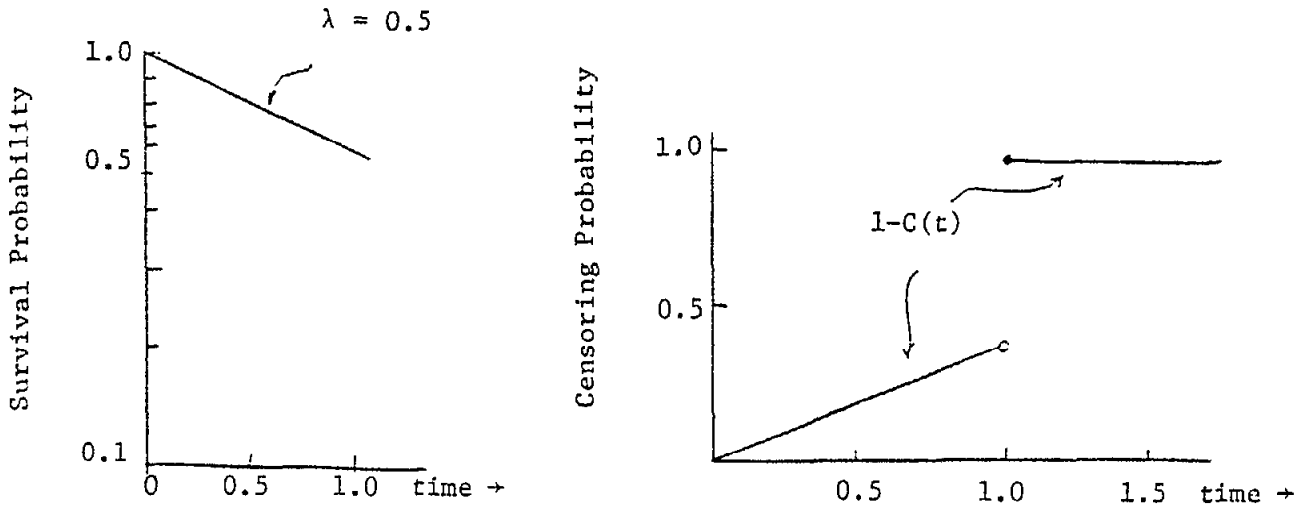


FIGURE 4.6: CONFIGURATION 6

Expected Percent Censored: 67.9%

Sample Size:	$N_1 = N_2 = 20$		$N_1 = N_2 = 50$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.008	.033	.006	.034
K^0	.001	.030	.005	.030
K^1	.003	.031	.007	.032
K^2	.003	.033	.007	.033
K^3	.003	.035	.009	.034
K^4	.003	.033	.009	.035
Gehan-Wilcoxon	.010	.052	.009	.038
Log-Rank	.010	.046	.007	.036

TABLE 4.7 POWER

Monte-Carlo Estimates of the Power of the Generalized Smirnov, K^{α} (α = 0,1,2,3,4) Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of H_0: $S_1 = S_2$ vs H_1: $S_1 < S_2$ (500 simulations)

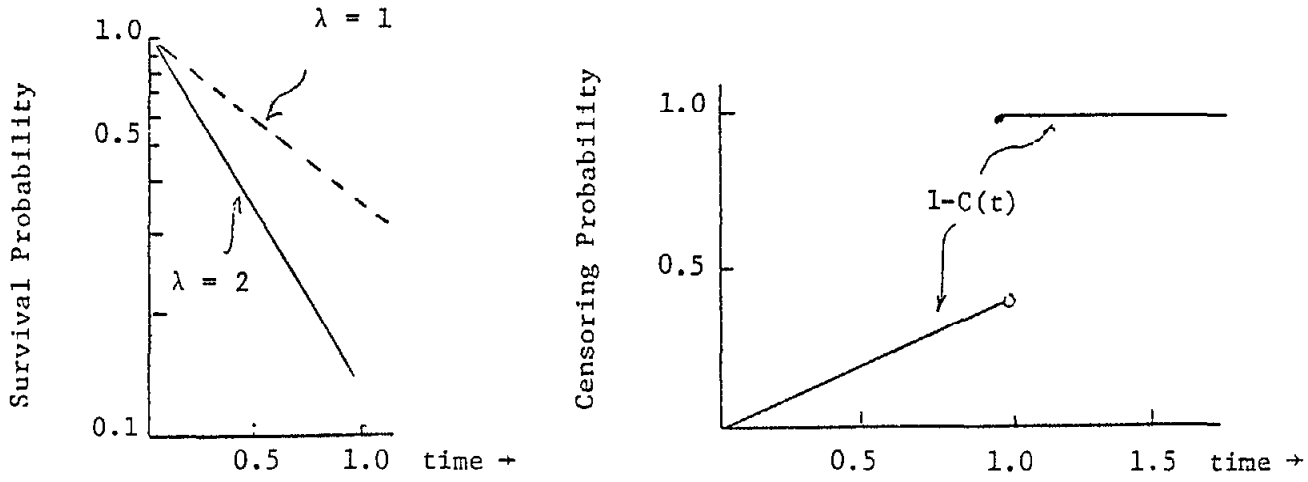


FIGURE 4.7: CONFIGURATION 7

Sample Size:	$N_1 = N_2 = 20$		$N_1 = N_2 = 50$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.236	.470	.566	.798
K^0	.068	.392	.410	.752
K^1	.156	.446	.570	.832
K^2	.166	.426	.542	.784
K^3	.140	.372	.466	.700
K^4	.110	.354	.396	.618
Gehan-Wilcoxon	.228	.478	.568	.820
Log-Rank	.276	.522	.640	.872

TABLE 4.8 POWER

Monte-Carlo Estimates of the Power of the Generalized Smirnov, K^{α} (α = 0,1,2,3,4) Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of H_0: $S_1 = S_2$ vs H_1: $S_1 < S_2$ (500 simulations)

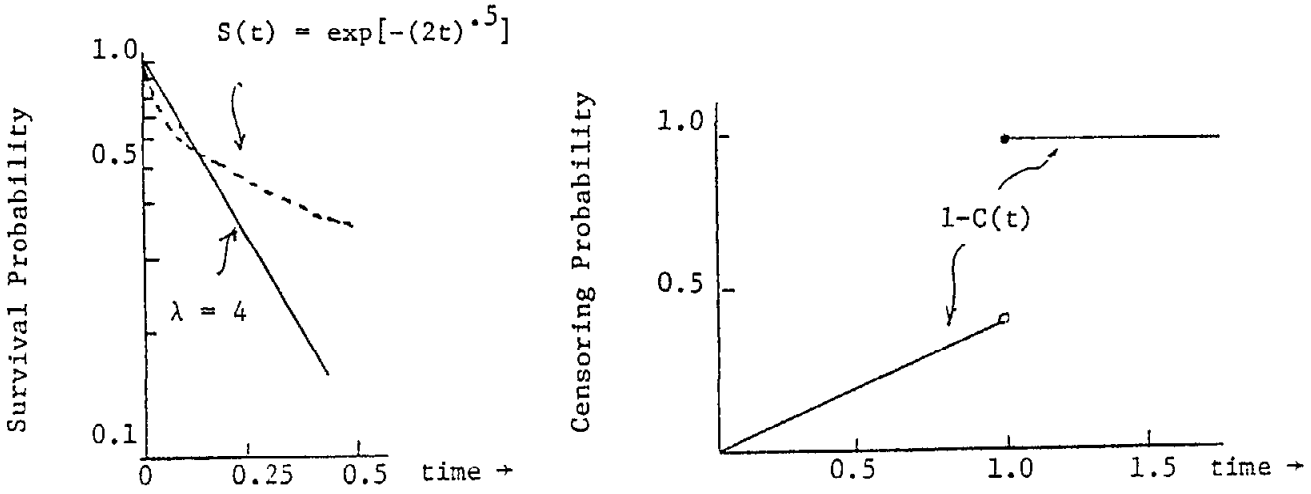


FIGURE 4.8: CONFIGURATION 8

Sample Size:	$N_1 = N_2 = 20$		$N_1 = N_2 = 50$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.252	.426	.638	.818
K^0	.046	.450	.578	.932
K^1	.092	.304	.414	.710
K^2	.034	.154	.136	.280
K^3	.012	.068	.024	.090
K^4	.010	.034	.002	.014
Gehan-Wilcoxon	.034	.130	.064	.188
Log-Rank	.140	.358	.418	.692

Monte-Carlo Estimates of the Power of the Generalized Smirnov, K^{α} (α = 0,1,2,3,4), Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of H_0: $S_1 = S_2$ vs H_1: $S_1 < S_2$ (500 simulations)

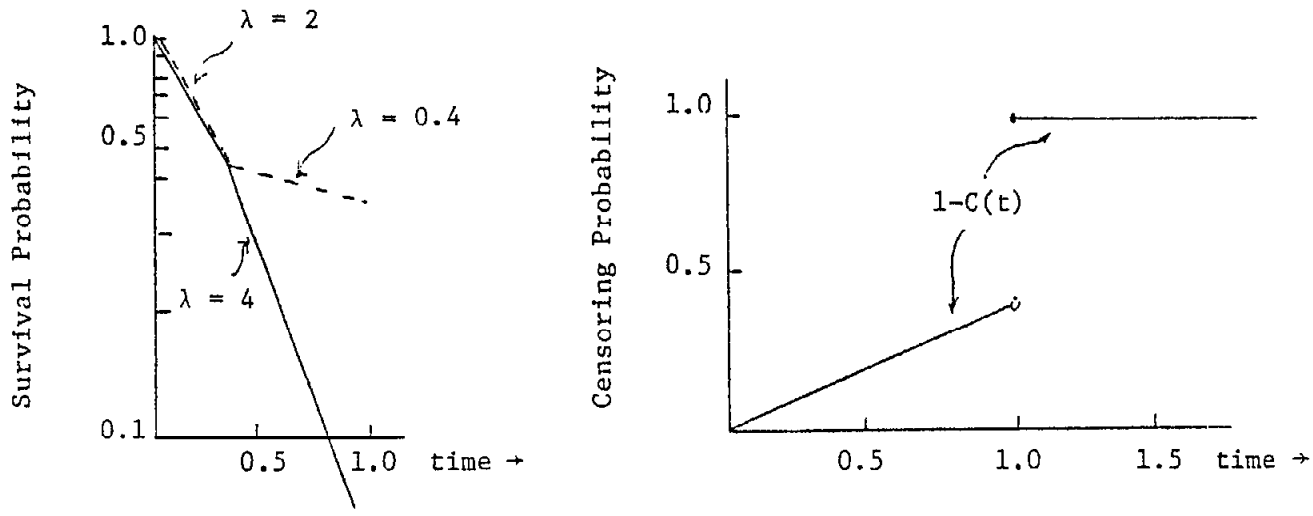


FIGURE 4.9: CONFIGURATION 9

Sample Size:	$N_1 = N_2 = 20$		$N_1 = N_2 = 50$	
Level of Test	.01	.05	.01	.05
Generalized Smirnov	.310	.488	.714	.854
K^0	.050	.552	.783	.948
K^1	.074	.364	.430	.724
K^2	.034	.196	.190	.348
K^3	.016	.100	.070	.192
K^4	.010	.054	.024	.098
Gehan-Wilcoxon	.028	.148	.080	.254
Log-Rank	.150	.416	.406	.688

TABLE 4.10 POWER

Monte-Carlo Estimates of the Power of the Generalized Smirnov, K^{α} ($\alpha = 0,1,2,3,4$), Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of $H_0: S_1 = S_2$ vs $H_1: S_1 < S_2$ (500 simulations)

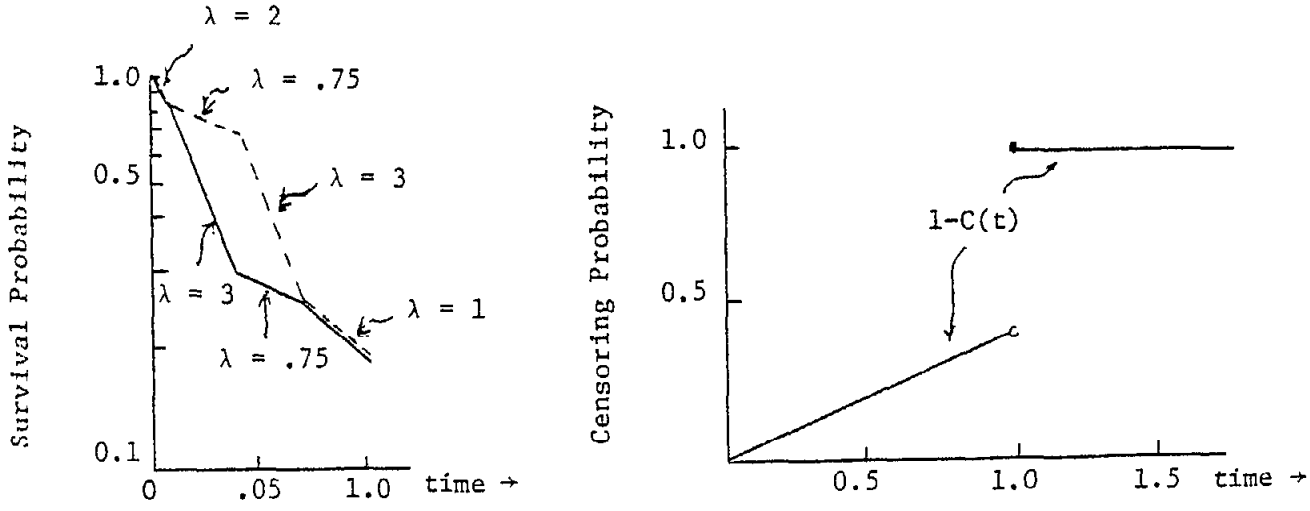


FIGURE 4.10: CONFIGURATION 10

Sample Size:	$N_1 = N_2 = 20$		$N_1 = N_2 = 50$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.198	.400	.604	.804
K^0	.010	.072	.060	.304
K^1	.052	.280	.346	.674
K^2	.084	.294	.386	.666
K^3	.080	.250	.302	.530
K^4	.068	.194	.202	.432
Gehan-Wilcoxon	.072	.274	.246	.504
Log-Rank	.054	.178	.132	.330

TABLE 4.11 POWER

Monte-Carlo Estimates of the Power of the Generalized Smirnov, K^{α} ($\alpha = 0,1,2,3,4$), Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of $H_0: S_1 = S_2$ vs $H_1: S_1 < S_2$ (500 simulations)

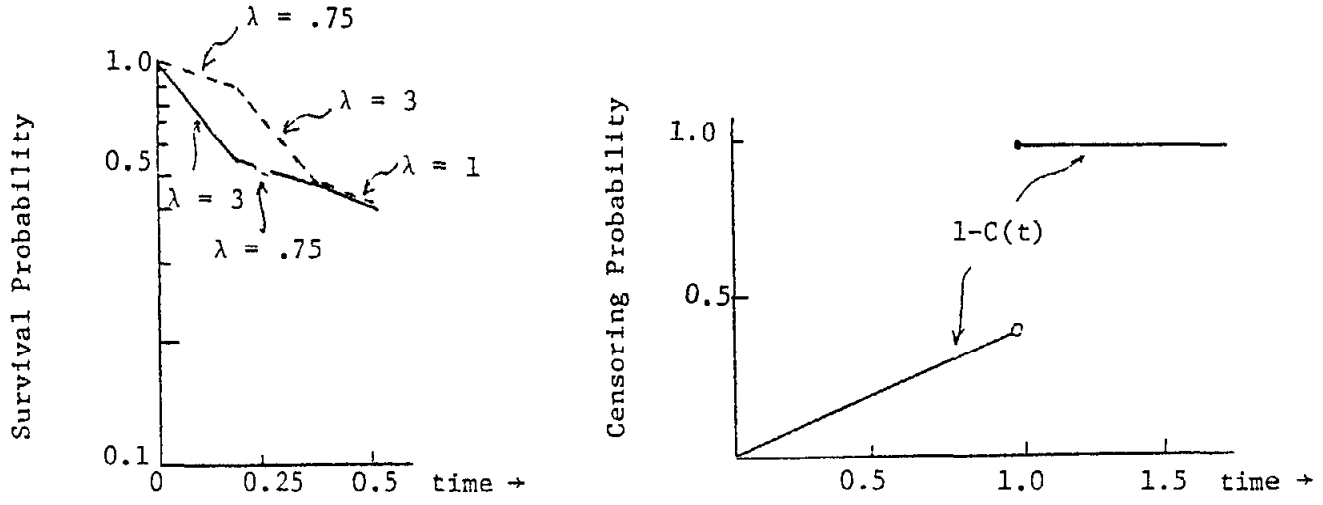


FIGURE 4.11: CONFIGURATION 11

Sample Size:	$N_1 = N_2 = 20$		$N_1 = N_2 = 50$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.144	.368	.580	.822
K^0	.000	.052	.012	.120
K^1	.032	.234	.276	.646
K^2	.108	.402	.580	.842
K^3	.162	.498	.700	.894
K^4	.186	.534	.732	.896
Gehan-Wilcoxon	.102	.322	.262	.522
Log-Rank	.042	.162	.068	.214

TABLE 4.12 POWER

Monte-Carlo Estimates of the Power of the Generalized Smirnov, K^{α} ($\alpha = 0,1,2,3,4$), Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of $H_0: S_1 = S_2$ vs $H_1: S_1 < S_2$ (500 simulations)

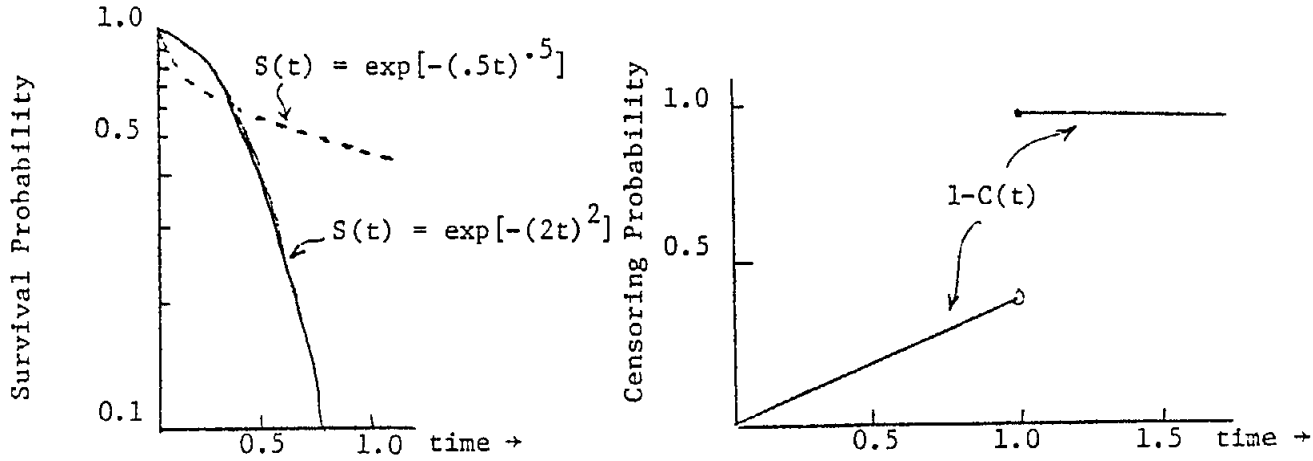
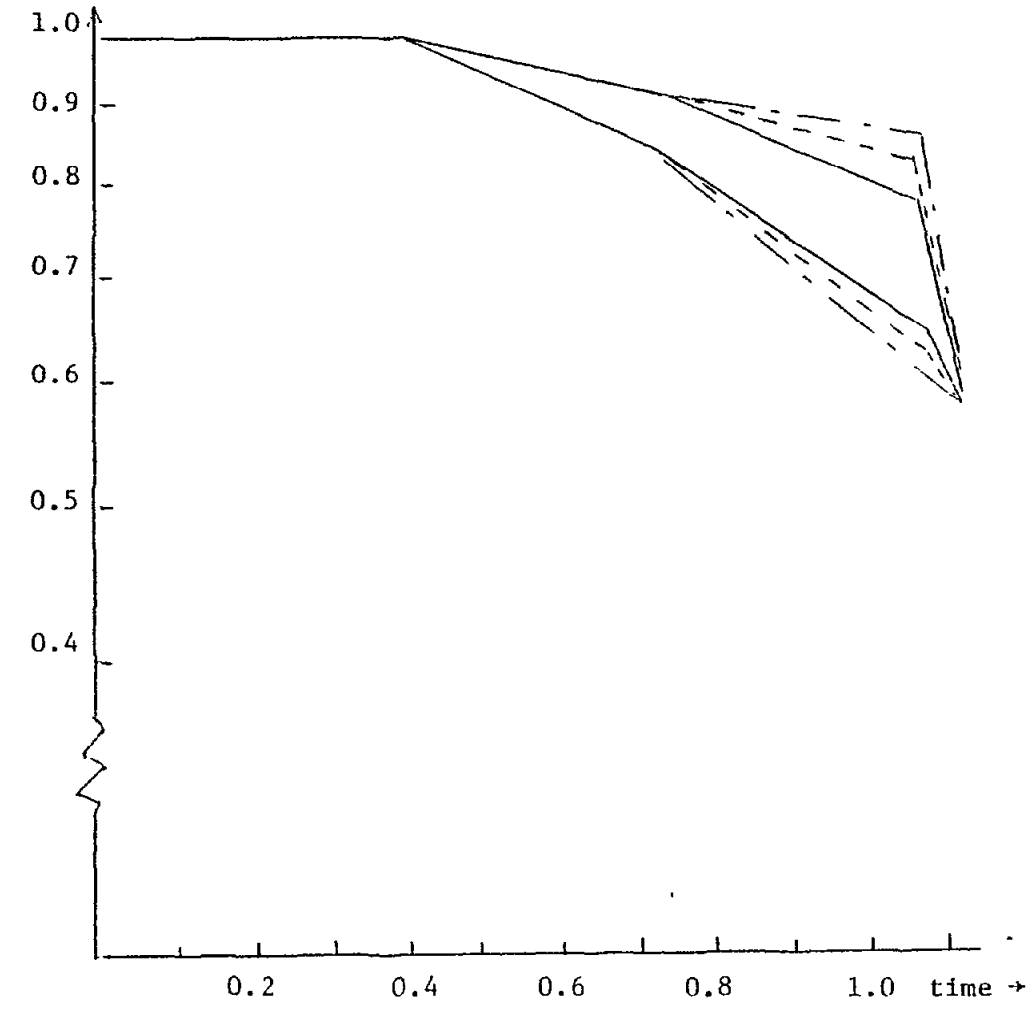
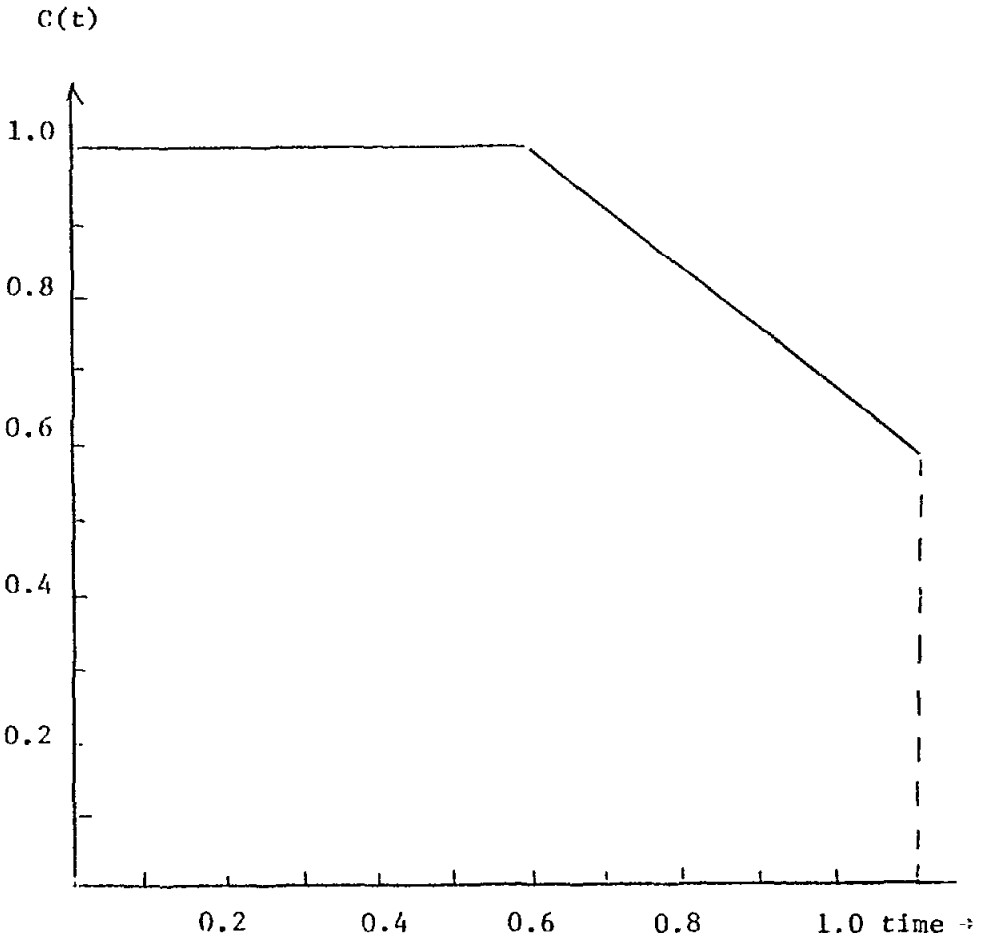


FIGURE 4.12: CONFIGURATION 12 .

Sample Size:	$N_1 = N_2 = 20$		$N_1 = N_2 = 50$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.680	.840	.992	.998
K^0	.264	.858	.990	1.000
K^1	.258	.642	.914	.980
K^2	.110	.362	.512	.760
K^3	.070	.202	.226	.406
K^4	.044	.110	.064	.168
Gehan-Wilcoxon	.068	.188	.154	.366
Log-Rank	.352	.656	.882	.974

FIGURE 4.13





Configuration 13 ————

Configuration 14 - - - - - -

Configuration 15 —— - ——

Note: See the following page for numerical values of hazard and survival functions.

INFORMATION FOR CONFIGURATIONS 13-15

Values of t: (t_a, t_b)	λ_1	λ_2	$S_1(t_b)$	$S_2(t_b)$
		CONFIGURATION 13		
(0.00, 0.36)	.03	.03	.99	.99
(0.36, 0.72)	.35	.18	.87	.93
(0.72, 1.08)	.74	.44	.67	.79
(1.08, 1.11)	4.70	8.70	.58	.61
		CONFIGURATION 14		
(0.00, 0.36)	.03	.03	.99	.99
(0.36, 0.72)	.35	.18	.87	.93
(0.72, 1.08)	.91	.30	.63	.83
(1.08, 1.11)	2.67	12.10	.58	.58
		CONFIGURATION 15		
(0.00, 0.36)	.03	.03	.99	.99
(0.36, 0.72)	.35	.18	.87	.93
(0.72, 1.08)	1.04	.17	.60	.87
(1.08, 1.11)	1.04	13.70	.58	.58

Monte-Carlo Estimates of the Power of the Generalized Smirnov, K^{α} ($\alpha = 0,1,2,3,4$), Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of $H_0: S_1 = S_2$ vs $H_1: S_1 < S_2$ (500 simulations)

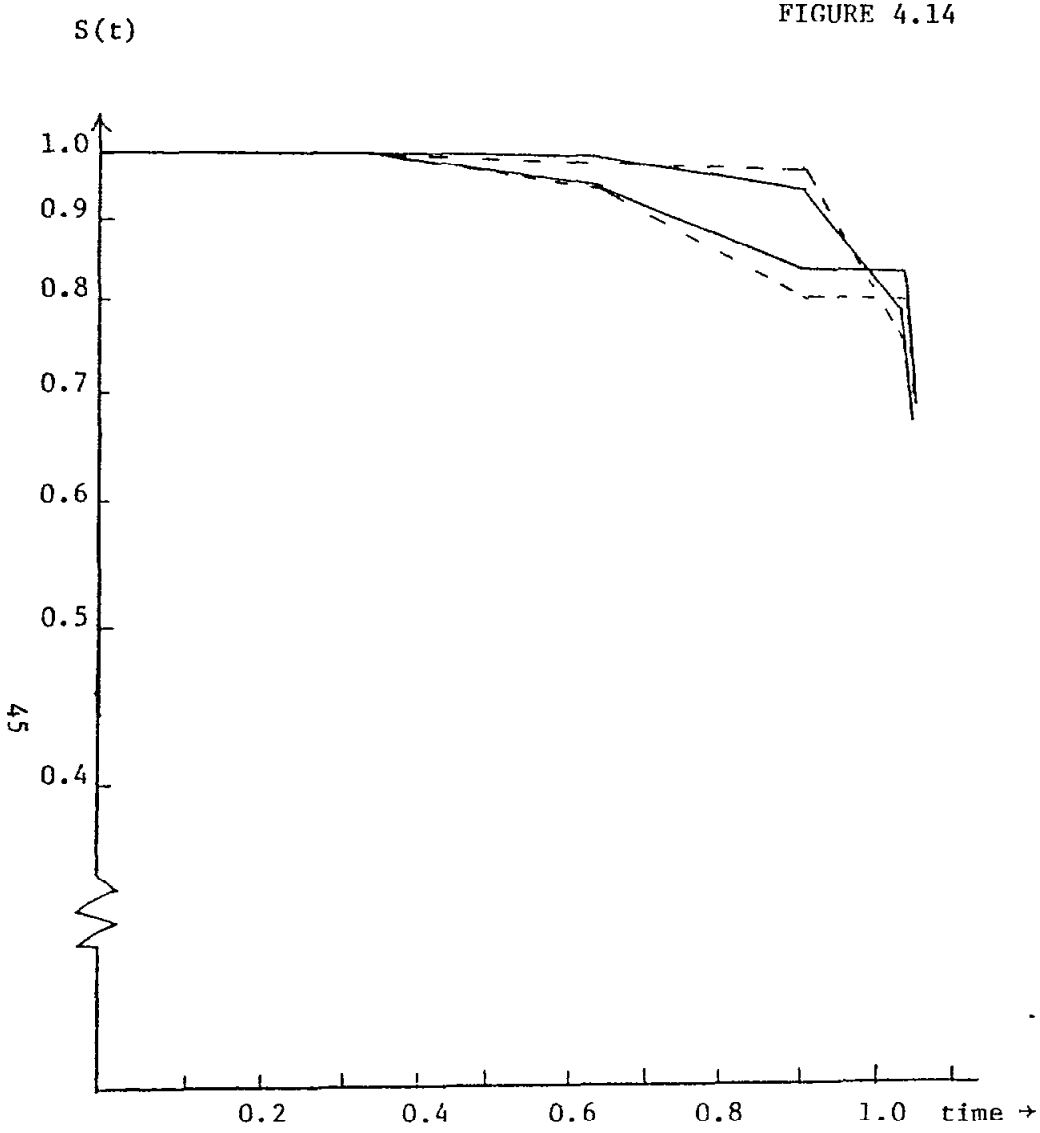
TABLE 4.13 (FOR CONFIGURATION 13)

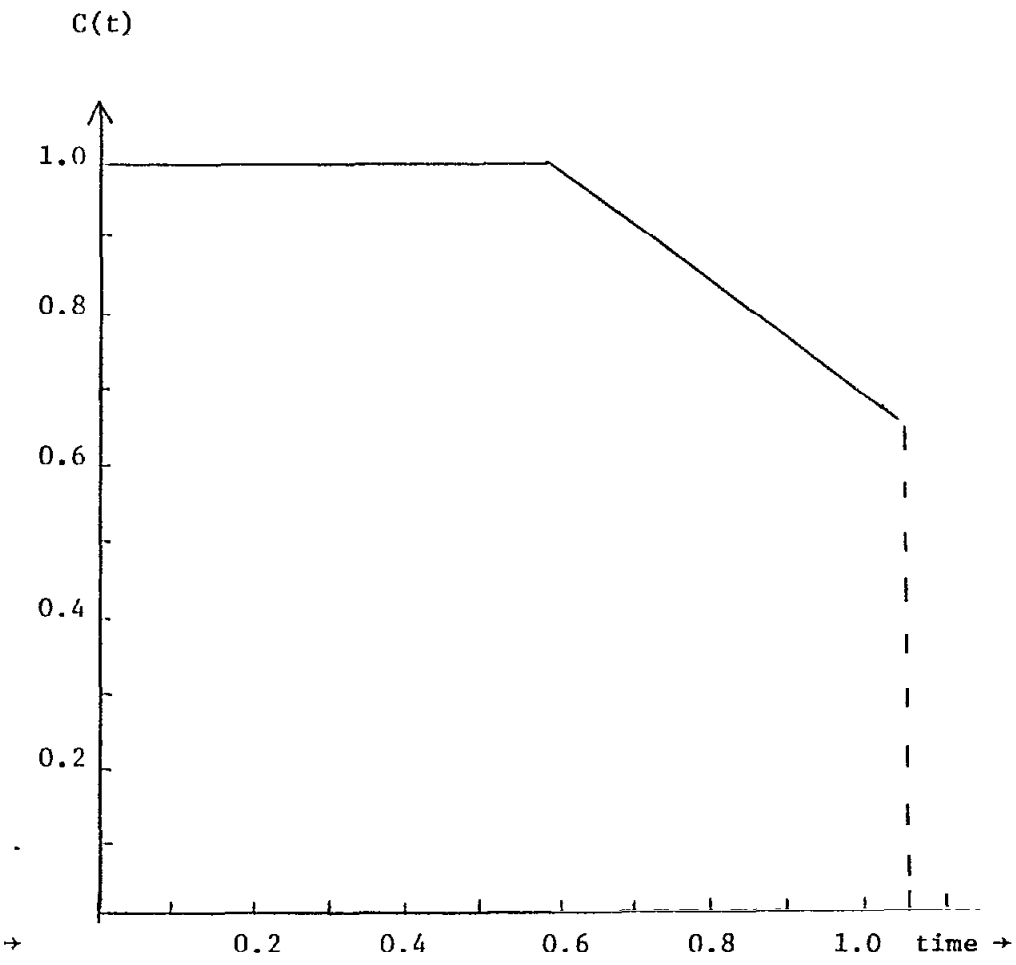
Sample Size:	$N_1 = N_2 = 50$		$N_1 = N_2 = 100$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.060	.200	.152	.400
K^0	.022	.144	.084	.302
K^1	.044	.180	.120	.394
K^2	.056	.208	.174	.442
K^3	.068	.234	.208	.478
K^4	.072	.238	.218	.478
Gehan-Wilcoxon	.064	.226	.160	.398
Log-Rank	.048	.154	.074	.232

TABLE 4.14 (FOR CONFIGURATION 14)

Generalized Smirnov	.154	.458	.494	.778
K^0	.054	.314	.274	.608
K^1	.112	.412	.408	.698
K^2	.144	.446	.484	.776
K^3	.168	.466	.504	.772
K^4	.172	.456	.498	.766
Gehan-Wilcoxon	.096	.316	.234	.524
Log-Rank	.044	.166	.094	.258

FIGURE 4.14





Configuration 16 ——

Configuration 17 - - - -

Note: See the following page for numerical values of hazard and survival functions.

Smirnov, K^{α} ($\alpha = 0,1,2,3,4$), Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of $H_0: S_1 = S_2$ vs $H_1: S_1 < S_2$ (500 simulations)

TABLE 4.15

Sample Size:	$N_1 = N_2 = 50$		$N_1 = N_2 = 100$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.362	.630	.810	.934
K^0	.166	.480	.618	.878
K^1	.262	.576	.734	.916
K^2	.336	.608	.774	.920
K^3	.362	.622	.774	.920
K^4	.362	.610	.758	.910
Gehan-Wilcoxon	.180	.406	.408	.658
Log-Rank	.068	.192	.140	.344

INFORMATION FOR CONFIGURATION 16-17

Values of t: (t_a, t_b)	λ_1	λ_2	$S_1(t_b)$	$S_2(t_b)$
		CONFIGURATION 16		
(0.00, 0.35)	.00	.00	1.00	1.00
(0.35, 0.65)	.20	.00	.94	1.00
(0.65, 0.91)	.49	.23	.83	.94
(0.91, 1.03)	.00	1.47	.83	.79
(1.03, 1.05)	9.10	7.50	.69	.68
		CONFIGURATION 17		
(0.00, 0.35)	.00	.00	1.00	1.00
(0.35, 0.65)	.20	.05	.94	.99
(0.65, 0.91)	.63	.05	.80	.97
(0.91, 1.03)	.00	2.06	.80	.76
(1.03, 1.05)	7.30	5.60	69	.68

Monte-Carlo Estimates of the Power of the Generalized Smirnov, K^{α} (α = 0,1,2,3,4), Gehan-Wilcoxon and Log-Rank One-Sided Test Procedures of H_0: $S_1 = S_2$ vs H_1: $S_1 < S_2$ (500 simulations)

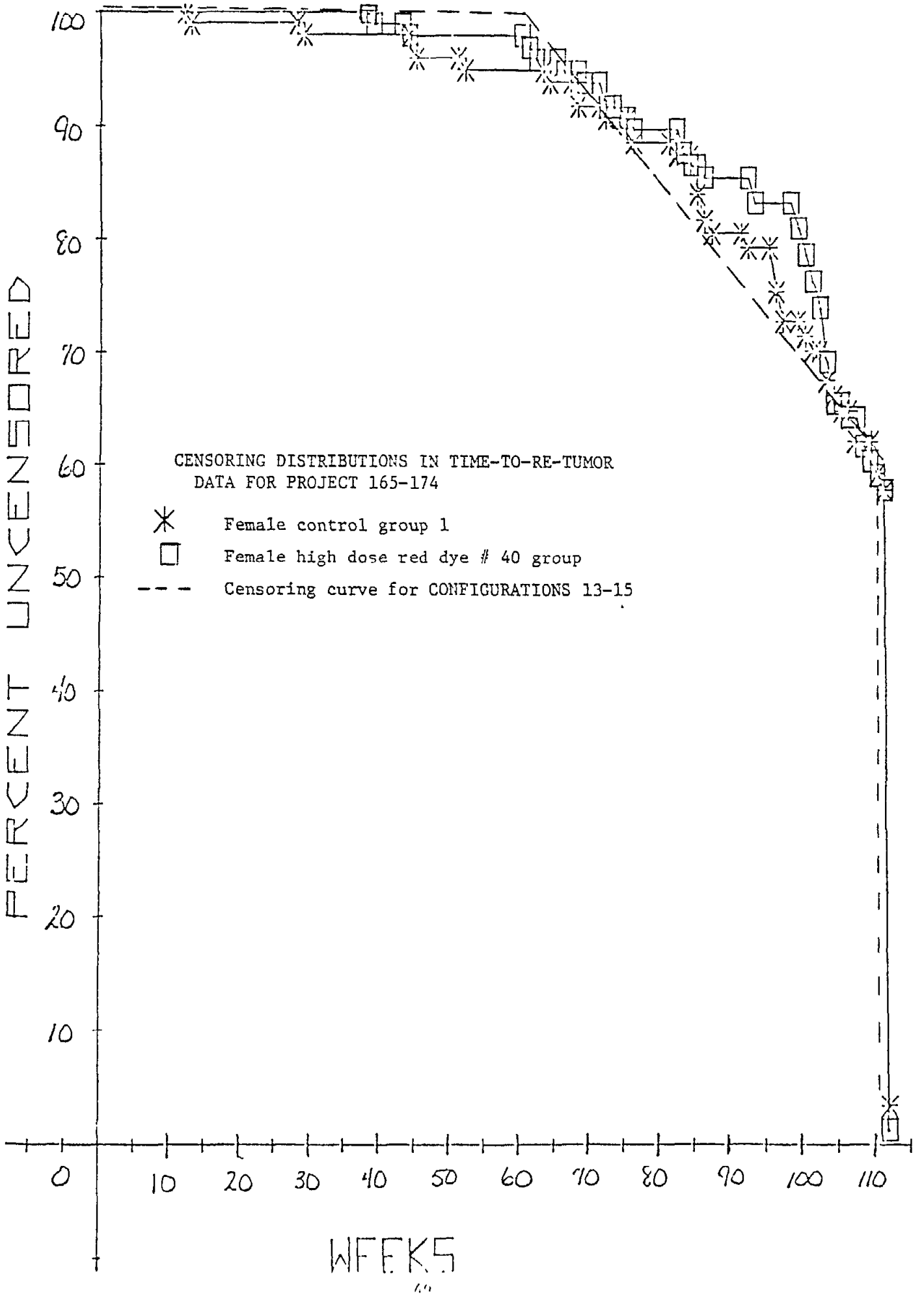
TABLE 4.16 (FOR CONFIGURATION 16)

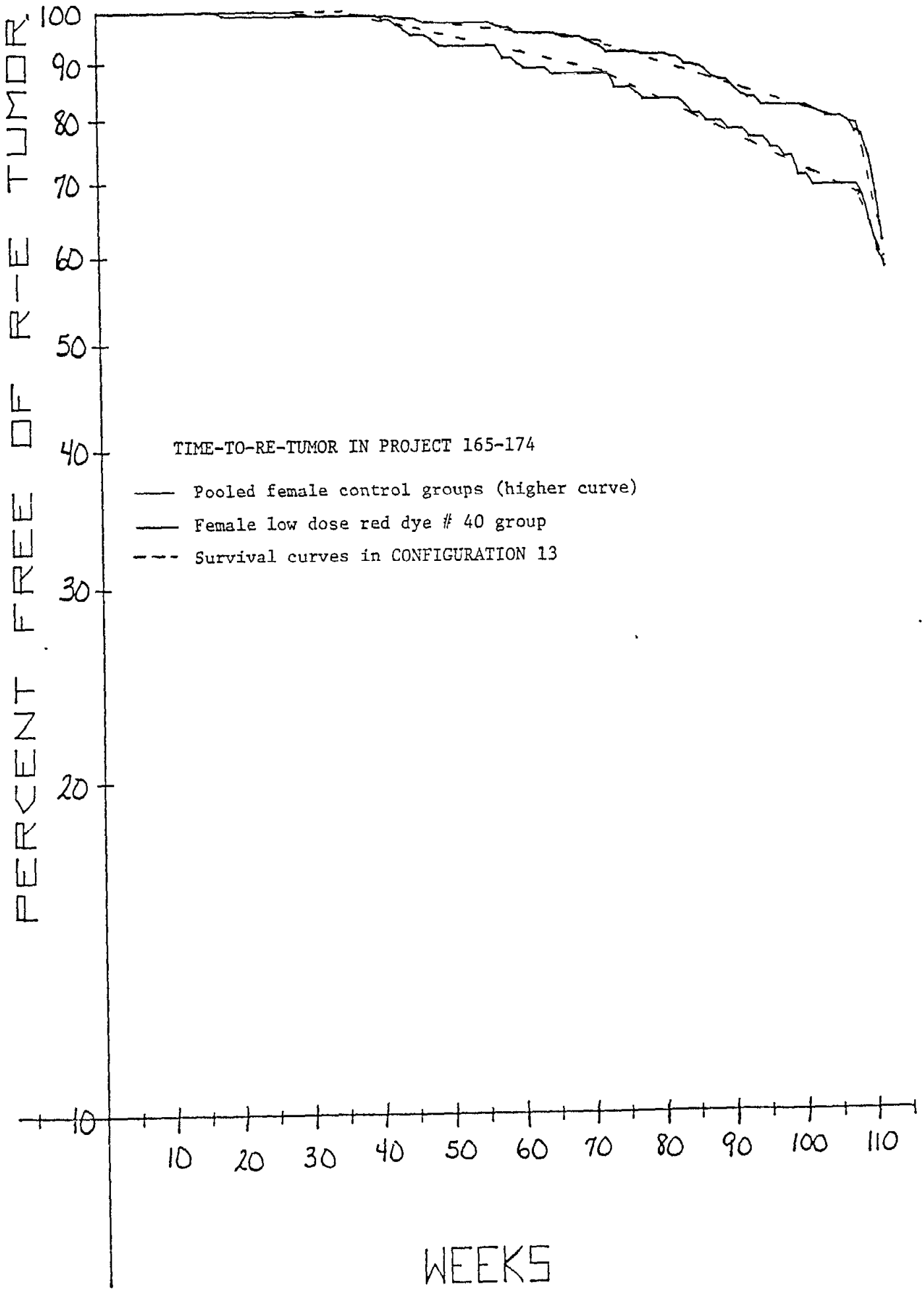
Sample Size:	$N_1 = N_2 = 50$		$N_1 = N_2 = 100$	
Level of Test:	.01	.05	.01	.05
Generalized Smirnov	.020	.144	.076	.322
K^0	.008	.080	.014	.142
K^1	.016	.126	.046	.262
K^2	.024	.182	.108	.402
K^3	.056	.236	.166	.524
K^4	.088	.302	.244	.620
Gehan-Wilcoxon	.040	.170	.044	.194
Log-Rank	.020	.090	.022	.072

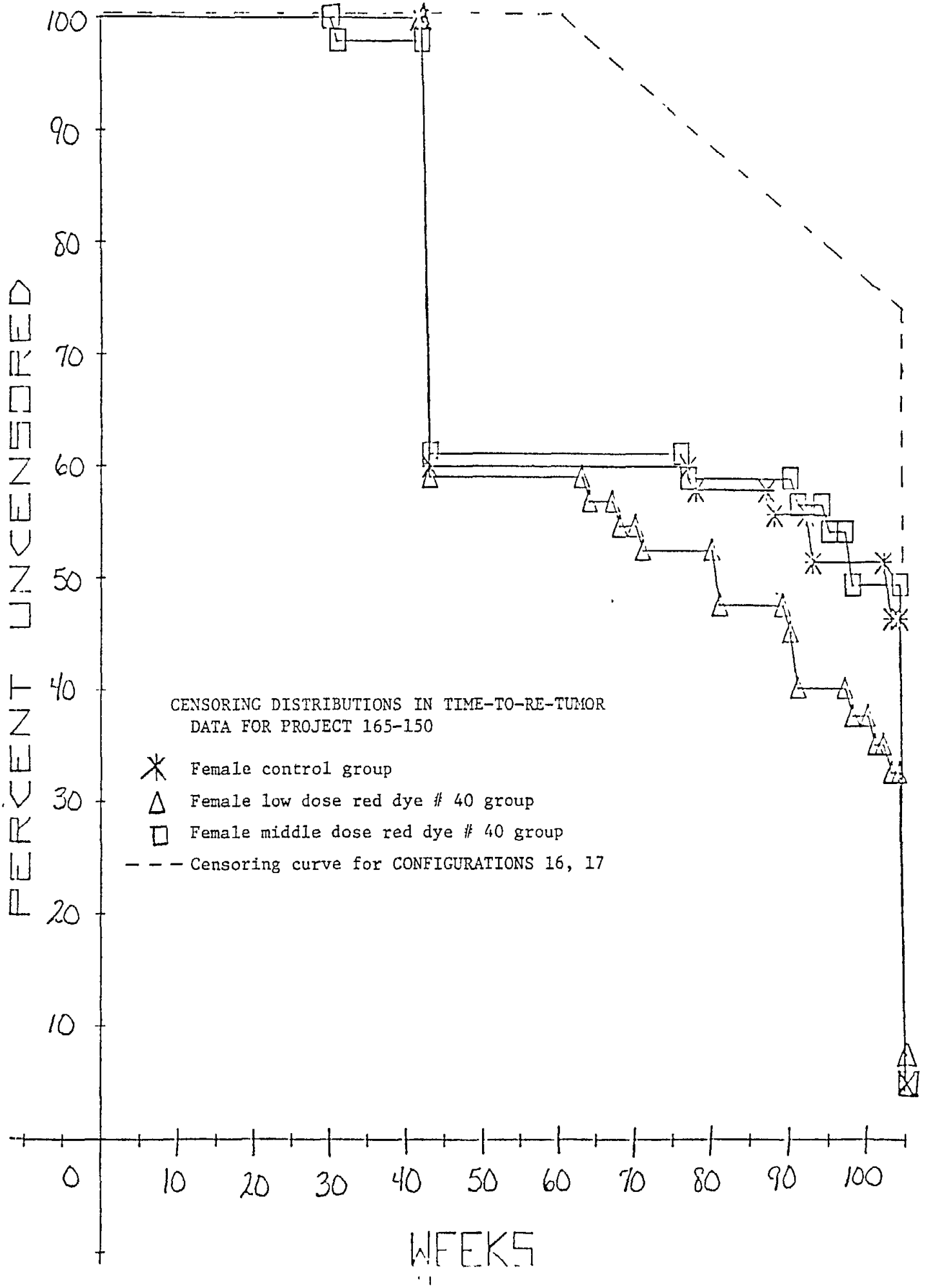
TABLE 4.17 (FOR CONFIGURATION 17)

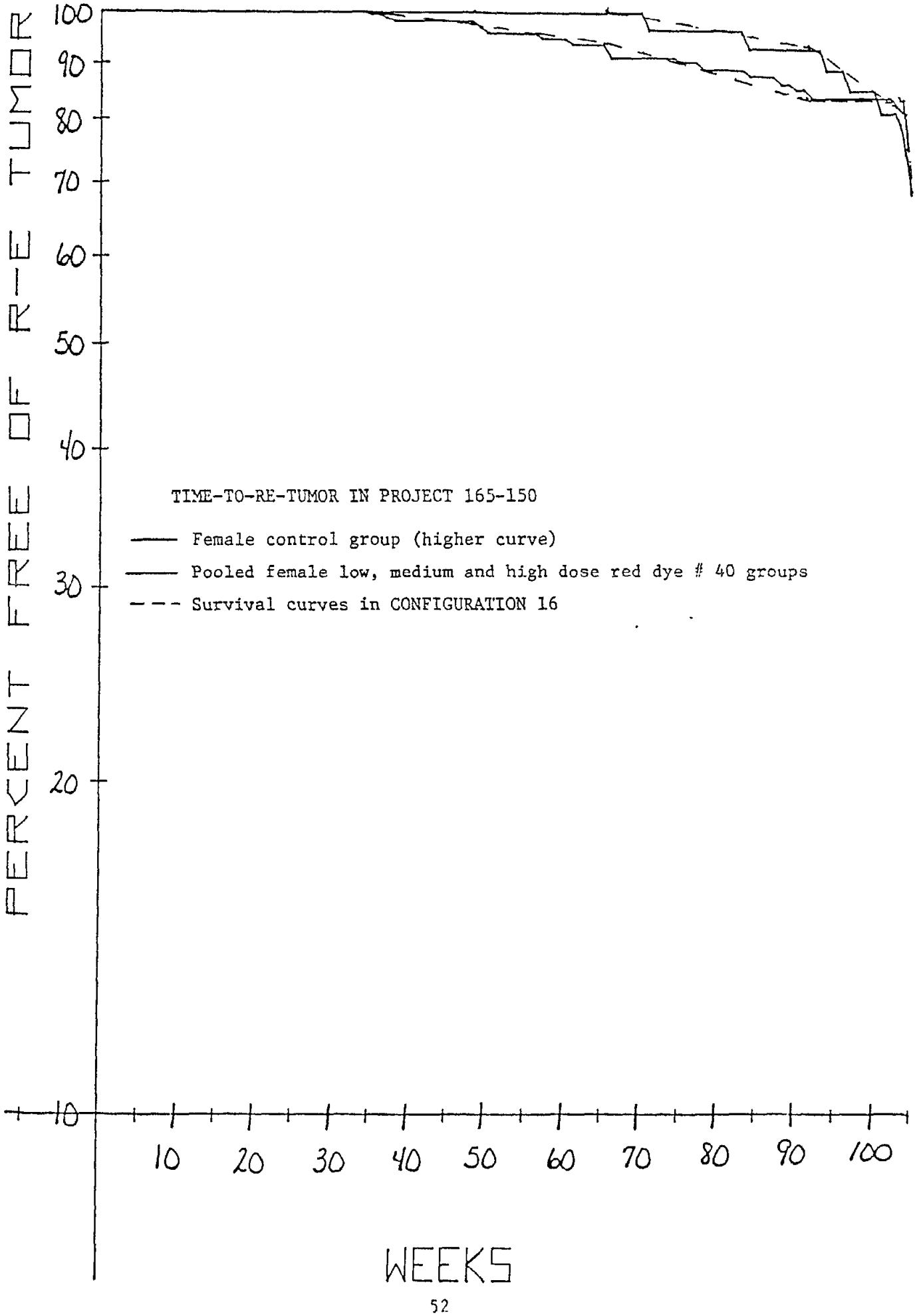
Generalized Smirnov	.060	.324	.460	.812
K^0	.012	.132	.146	.552
K^1	.040	.254	.298	.734
K^2	.074	.364	.482	.830
K^3	.118	.450	.596	.886
K^4	.166	.526	.688	.916
Gehan-Wilcoxon	.046	.174	.112	.332
Log-Rank	.016	.074	.032	.122

FIGURE 4.15


CENSORING DISTRIBUTIONS IN TIME-TO-RE-TUMOR DATA FOR PROJECT 165-174
Female control group 1
Female high dose red dye # 40 group
Censoring curve for CONFIGURATIONS 13-15
PERCENT UNCENSORED
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
60
70
80
90
100
110
WEEKS


PERCENT FREE OF RE-TUMOR
100
90
80
70
60
50
40
30
20
10
10
20
30
40
50
60
70
80
90
100
110
WEEKS
TIME-TO-RE-TUMOR IN PROJECT 165-174
Pooled female control groups (higher curve)
Female low dose red dye # 40 group
Survival curves in CONFIGURATION 13


PERCENT UNCENSORED
100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
60
70
80
90
100
WEEKS
CENSORING DISTRIBUTIONS IN TIME-TO-RE-TUMOR
DATA FOR PROJECT 165-150
Female control group
Female low dose red dye # 40 group
Female middle dose red dye # 40 group
Censoring curve for CONFIGURATIONS 16, 17


PERCENT FREE OF RE-TUMOR
100
90
80
70
60
50
40
30
20
10
10
20
30
40
50
60
70
80
90
100
WEEKS
TIME-TO-RE-TUMOR IN PROJECT 165-150
Female control group (higher curve)
Pooled female low, medium and high dose red dye # 40 groups
Survival curves in CONFIGURATION 16

V. References

Fleming, T. R. and Harrington, D. P. (1979). A class of hypothesis tests for one and two samples of censored survival data. Submitted to Communications in Statistics.

Fleming, T. R., O'Fallon, J. R., O'Brien, P. C., and Harrington, D. P. (1979). Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right censored data. Submitted to Biometrics.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). Journal of the Royal Statistical Society, Series A 135, 185-207.

Prentice, R. L., and Marek, P. (1979). A qualitative discrepancy between censored data rank tests. Biometrics, to appear.

Knuth, D. E. (1969). The Art of Computer Programming: Volume 2, Seminumerical Algorithms. Addison-Wesley, Reading, Massachusetts.