SOME NONPARAMETRIC TWO SAMPLE TESTS WITH RANDOMLY CENSORED DATA

MAI ZHOU

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY
1986

ABSTRACT

SOME NONPARAMETRIC TWO SAMPLE TESTS WITH RANDOMLY CENSORED DATA

MAI ZHOU

In this thesis, we propose and study a new generalization of the two sample rank tests of Chernoff and Savage (1958) by replacing the usual empirical distribution function by the Kaplan-Meier estimator. This new class of rank tests which can accommodate randomly right censored data is not the same as Gill's K-class of tests. Its asymptotic distribution and efficacy are derived and studied. We show how to choose the optimal test within the class and prove that it is the most powerful if the censoring patterns are the same. A consistent null variance estimator of the test statistic is given in Chapter 6.

We also study several versions of a difference of means test for censored data. Recently developed censoring regression techniques are specialized to the two sample case. Some interesting results are obtained when we compare their asymptotic variances.

Some Monte Carlo simulation comparisons are carried out for a sample size of 50 with the well known Mantel-Haenszel or log-rank test included as a standard one. The resulting tables are given in Chapter 8.

Some interesting new properties of the Kaplan-Meier estimator are also derived. These results are summerized in Chapter 2.

Table of Contents

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Professor John Van Ryzin, for his constant encouragement, patience and invaluable guidance. Without his help this thesis would not have been written.

I am also grateful to Professor Tze Leung Lai for the help he gave me during the preparation of this thesis.

Thanks are also due to the other members of the thesis defense committee, Professor Howard Levene, Professor Bruce Levin, and Professor Zvi Schechner for their helpful comments.

I also want to thank Dr. K. K. Gordon Lan for some helpful discussions.

The financial support for this research was provided in part by National Institute of Health grants No. GM 28405 and No. GN 27976, and by a fellowship from the Department of Statistics at Columbia University.

To My Parents.

CHAPTER 1

INTRODUCTION

In many medically related statistical problems in which a lifetime or time to occurrence of some event is observed, the data are typically nonnegative and often incomplete due to a variety of reasons. For instance, during the observation period a patient may not want to continue to use the drug or may die due to competing causes not under study. Hence, we sometimes only observe a portion of the lifetime, and such data are said to be censored. The analysis of such censored survival data has been given much attention in the recent literature. As some examples consider the following statistical problems:

Example 1 (Kaplan-Meier estimator)

Suppose the actual lifetimes of the patients are $X_1 , X_2 , \cdots , X_n$ which are i.i.d. with $P(X_i < t) = F(t)$. Unfortunately, we can only observe the pair (Z_i , δ_i) , where

$$Z_i = min(X_i , C_i) ; \qquad \delta_i = I_{[X_i \leq C_i]} \ .$$

Here $C_1 , C_2 , \cdots , C_n$ are called censoring or follow-up times. We make the added assumption that the C_i's are i.i.d. with $P(C_i < t) = G(t)$ and that they are independent of the X_i's. This is the so called random censorship model (see, e.g., Breslow and Crowley (1972)).

From the observations (Z_i , δ_i) we wish to estimate the distribution $F(t)$ without assuming any parametric structure on F ; that is, we want a nonparametric estimator of F.

Kaplan and Meier (1958) suggest the so-called product limit estimator , defined as follows,

$$\hat{F}(t) = 1 - \prod_{s \le t}(1 - \frac{\Delta N(s)}{Y(s)}) \tag{1.1}$$

to estimate $F(t)$, where

$$N(t) = \#\{\ i : Z_i \le t\ ;\ \delta_i = 1\ \} \qquad \text{and} \qquad Y(t) = \#\{\ i : Z_i \ge t\ \}\ ,$$

i.e., $N(t)$ is the number of persons who died up to time t, $Y(t)$ is the number of people at risk at time t, and $\Delta N(s)$ represents the jump size of $N(t)$ at s (i.e., the number of subjects dying at s). The set $\{\ i : Z_i \ge t\ \}$ is called the risk set at time t.

Since being introduced, various good properties of the Kaplan-Meier estimator have been studied. These include strong, uniform convergence with and without rate; and asymptotic normality at a fixed point and as a process. (See Breslow and Crowley (1972); Peterson (1977); Phadia and Van Ryzin (1978); Foldes and Rejto (1981) and Gill (1983)).

Example 2 (Two sample rank tests)

In medical studies, industrial life-testing, and in many other situations, we often want to compare the survival distributions of two populations. For example, there is a new drug which we want to compare with the 'standard' one, or we might want to see whether there is any difference between the two treatments (or two products) in terms of lifetime.

Naturally the data coming from the two experiments fall into two groups which are subject to possibly different censorings. The aim is to make statistical inference on the lifetime distributions regardless of the censoring and if possible to use a nonparametric procedure which is robust.

Specifically, we have two sets of data, n_1 of one kind, n_2 of the other and each group is subject to censoring. Let X_{i1} , X_{i2} , $\cdots$, X_{in_1} be i.i.d. lifetimes for i=1,2, $C_{i1}, C_{i2}, \cdots, C_{in_1}$ be i.i.d. censoring times, and $P(X_{ij} < t) = F_i(t)$; $P(C_{ij} < t) = G_i(t)$. However, we only observe (Z_{ij}, δ_{ij}), where $Z_{ij} = \min (X_{ij}, C_{ij})$, $\delta_{ij} = I_{[X_{ij} \le C_{ij}]}$. The task is to test if there is any difference between F_1 and F_2.

The Mantel-Haenszel test, perhaps the most widely used test today, was proposed by Mantel and Haenszel (1959), Mantel (1966), Peto and Peto (1972) and by Cox (1972) among others. Thus it sometimes is called the log-rank test or the Cox test.

The Mantel-Haenszel test uses the test statistic

$$\int_0^\infty \frac{Y_1 Y_2}{Y_1 + Y_2} \left(\frac{dN_1}{Y_1} - \frac{dN_2}{Y_2} \right) \tag{1.2}$$

where $Y_i ; N_i$ bear the same meaning as in the Example 1 except that the subscript which indicates the process is associated with the sample i .

Other two sample tests can be written in the similar way. For example, Gehan's test (Gehan (1965)) which is a censored two sample generalization of the Mann-Whitney-Wilcoxon test; the Tarone-Ware class of statistics (Tarone and Ware (1977)) and the Harrington-Fleming G_ρ statistics (Harrington and Fleming (1982)) can all be written similar to (1.2) with weight functions differing from $\frac{Y_1 Y_2}{Y_1 + Y_2}$ therein. In particular, all of these classes of statistics are special subclasses of the tests given by Gill (1980) referred to as his K-class for which he proved many optimal properties.

In this thesis, we will introduce and study (1) a new class of rank tests and (2) a difference of means tests which are seen to be alternatives for the two sample censored data problem. We now close this chapter with a thesis outline.

In Chapter two we present some new properties of the Kaplan-Meier estimator which will be used in later chapters. The extensions given therein shed new light on the Kaplan-Meier estimator.

Starting from Chapter three, we confine ourselves to the two sample problem. We propose a whole new class of rank statistics of censored data for detecting any difference in the F_i and prove its asymptotic normality in Chapter three. The proof of the negligibility of some of the higher order terms is postponed to Chapter five which is a long and technical one. This class is a generalization to censored data of the famous Chernoff-Savage (1958) tests for the uncensored data case.

Chapter four deals with the problem of choosing the best test (in the sense of maximizing Pitman efficacy) out of the new class for a specified setting. We give the defining equation which determines the best choice of the test in terms of the underlying null distribution, the censoring distributions, and the nature of the alternative hypotheses.

In Chapter six we propose a null variance estimator for the statistics proposed in Chapter three. We prove its consistency as an estimator. This enables us to perform the test by using the asymptotic theory based only on the observed data.

Chapter seven introduces and studies a mean test which is an analog of the t-test in the censorship case. We also obtain some interesting results when we apply several regression techniques to the two sample situation.

Some simulation comparison is done on a Masscomp MC500DP computer. The resulting tables as well as some theoretical comparisons will be presented in Chapter eight. Chapter nine is a summary chapter.

CHAPTER 2

SOME ONE SAMPLE RESULTS

In this chapter, we establish some one sample properties of the Kaplan-Meier estimator that will be needed later. Although we sometimes will only use the results in a special case, we think it is worthwhile to present them in a more general setting because they are of interest in their own right.

Suppose, throughout this chapter, that the nonnegative r.v.'s $X_1, X_2, \cdots, X_n$ are i.i.d. lifetimes with $P(X_1 \leq t) = F(t)$; $C_1, C_2, \cdots, C_n$ are i.i.d. censoring times with $P(C_1 \leq t) = G(t)$, both *F(t)* and *G(t)* are not necessarily continuous, and X_i's are independent of C_i's.

Suppose we cannot observe all the X_i's but only $Z_i = X_i \wedge C_i$ and $\delta_i = I_{[X_i \leq C_i]}$. A classical nonparametric estimator of *F(t)* based on (Z_i, δ_i)'s is the Kaplan-Meier or product limit estimator (see (1.1)). We will denote it by $\hat{F}_n(t)$ or sometimes simply $\hat{F}$.

Theorem 2.1 If $\theta(t)$ is a nonnegative function with $E\theta(X_i) < \infty$ and $\int_0^\infty I_{[\theta(t) \text{ is a discontinuity point }]} dF(t) = 0$, then

$$0 \leq \int_0^\infty \theta(t) dF(t) - E\int_0^\infty \theta(t) d\hat{F}_n(t) \leq \int_0^\infty [P(Z_1 \leq t)]^n \, \theta(t) dF(t) \quad . \tag{2.1}$$

Remark 2.1: Mauro (1985) derived the left half of the above inequality by assuming *F* and *G* are continuous and by using a combinatoric argument, while we allow both *F* and *G* to be discontinuous. Furthermore, we get the much harder right hand inequality. For instance, if we take $\theta(t) = I_{[t \leq u]}$, we get a bound on the bias of the Kaplan-Meier estimator, which improves upon Gill's (1980)[3.2.17] bound.

Let us first consider an application of the Theorem 2.1 before giving the proof of Theorem 2.1. Define, for any distribution function *F(t)* , $\tau_F = sup\ \{ t : F(t) < 1 \}$.

Theorem 2.2 If $\theta(t)$ is a real continuous function such that $E\theta(X_1)$ exists, and $\tau_G \geq \tau_F$, then we have

$$\int_0^\infty \theta(t) d\hat{F}_n(t) \xrightarrow{P} \int_0^\infty \theta(t) dF(t)$$

Furthermore, if $E\theta(X_1)$ is finite, then

$$\int_0^\infty \theta(t) d\hat{F}_n(t) \xrightarrow{L_1} \int_0^\infty \theta(t) dF(t) \qquad as\ n \to \infty.$$

Here $\xrightarrow{P}$ and $\xrightarrow{L_1}$ denote convergence in probability and in L_1-norm, respectively.

Proof of Theorem 2.2: If $E\theta(X_1)$ is finite, then the same argument as in Mauro (1985) [Theorem 4.1] shows $\int \theta d\hat{F} \xrightarrow{P} \int \theta dF = E\theta(X_1)$. If, however, $E\theta(X_1)$ is infinite, say $E\theta(X_1) = +\infty$, we define, for $K > 0$, $\theta^K(t) = min(\theta(t), K\ )$, then $E\theta^K(X_1) < \infty$ and

$$\int_0^\infty \theta(t) d\hat{F}_n(t) \geq \int_0^\infty \theta^K(t) d\hat{F}_n(t) \xrightarrow{P} E\theta^K(X_1) \quad as\ n \to \infty\ .$$

On the other hand, since $E\theta^K(X_1) \to E\theta(X_1) = +\infty$ as $K \to +\infty$, we must have $\int_0^\infty \theta(t) d\hat{F}(t) \xrightarrow{P} +\infty = E\theta(X_1)$. The case of $E\theta(X_1) = -\infty$ is similar.

For the L_1 convergence, notice that Theorem 2.1 implies, for nonnegative θ, $E\int \theta d\hat{F} \to E\theta(X_1)$. This fact together with the already shown weak convergence property implies (see Chow and Teicher, 1978, p. 100), $\int \theta d\hat{F} \xrightarrow{L_1} \theta(X_1)$.

By writing $\theta(t) = \theta^{+}(t) - \theta^{-}(t)$, where θ^{+} and θ^{-} are the positive and negative part of θ respectively, the L_1 convergence for a general θ follows easily. ■

We now present a series of lemmas followed by the proof of Theorem 2.1.

We first introduce some notation. Let

$$\Lambda(t) = \int_{[0,t]} \frac{dF(s)}{1-F(s-)} \ ;$$

$$N(t) = \#\{\ i : Z_i \le t,\ \delta_i = 1\ \}\ ;$$

$$Y(t) = \#\{\ i : Z_i \ge t\ \} \quad \text{and} \quad T = T_n = \max\ \{\ Z_1, Z_2, \ \cdots, Z_n\ \}\ .$$

It is well known that

$$M_i(t) = I_{[Z_i \le t,\ \delta_i = 1]} - \int_{[0,\ t]} I_{[Z_i \ge u]} d\Lambda(u)$$

are martingales (in fact square integrable martingales) for an appropriate choice of F_t ; in fact, the F_t can be chosen to be

$$F_s = \sigma\{\ \textit{all events that might be observed on}\ [0,\ s]\ \}$$

(a so called self-exciting filtration). Similarly

$$M(t) = \sum_1^n M_i(t) = N(t) - \int_{[0,t]} Y(t) d\Lambda(t)$$

is also a martingale with respect to F_t (see, e.g., Aalen (1978)).

Lemma 2.1

$$E\left[I_{[T_n > t]}(M_i(s+h) - M_i(s)\ )|F_s\right] \le 0 \quad a.s.$$

for $s < t$, $h > 0$ and $s + h < t$.

Proof: Consider, first, the set $\{Z_i < s\ \}$. On this set we can easily check $M_i(s + h) - M_i(s) = 0$. So by writing

$$E\{I_{[T_n > t\ ]}(M_i(s + h) - M_i(s))|F_s\} = E\{(I_{[Z_i < s\ ]} + I_{[Z_i \ge s\ ]})I_{[T_n > t\ ]}(M_i(s + h) - M_i(s))|F_s\}$$

$$= E\{I_{[Z_i < s]}I_{[T_n > t]}(M_i(s+h) - M_i(s))|F_s\} + E\{I_{[Z_i \geq s]}I_{[T_n > t]}(M_i(s+h) - M_i(s))|F_s\}$$

$$= I_{[Z_i < s]}E\{0\,|F_s\} + I_{[Z_i \geq s]}E\{I_{[Z_i \geq s]}I_{[T_n > t]}(M_i(s+h) - M_i(s))|F_s\} \ , \qquad (2.2)$$

we see immediately that

$$E[I_{[T_n > t]}(M_i(s+h) - M_i(s))|F_s] = 0 \qquad on\ \{Z_i < s\}\ .$$

What remains to be proved is that the second term of (2.2) above ≤ 0 almost surely.

Recall that

$$M_i(s) = I_{[\delta_i = 1,\, Z_i \leq s]} - \int_0^s I_{[Z_i \geq u]}d\Lambda(u) \quad .$$

Note, we have that

$$E[I_{[T_n > t]}\,(M_i(s+h) - M_i(s)\,)\mid F_s] = \qquad (2.3)$$

$$= E[I_{[T_n > t]}\,I_{[\delta_i = 1,\, s < Z_i \leq s+h]} \mid F_s] - E[I_{[T_n > t]}\int_s^{s+h} I_{[Z_i \geq u]}d\Lambda(u)\ |F_s]$$

$$= \int_{(s,\, s+h]} P(\delta_i = 1,\, T_n > t,\, Z_i \in dl \mid F_s) - E[I_{[T_n > t]}\,(\Lambda(Z_i \wedge s{+}h) - \Lambda(Z_i \wedge s)) \mid F_s]\ .$$

Without loss of generality, let us compute the conditional expectation on the set where

$$\{Z_i \geq s,\, Z^{j_1} \geq s,\, Z^{j_2} \geq s,\ \cdots\ ,\, Z^{j_k} \geq s,\, Z^{j_{k+1}} < s,\ \cdots\ ,\, Z^{j_{n-k-1}} < s\}\ ,$$

where $(k = 0,\, 1,\ \cdots\ ,\, n-1\,)$.

On this set, for $s < l < t$

$$P(\delta_i = 1,\, Z_i \in dl,\, T_n > t|F_s) = P(\delta_i = 1,\, Z_i \in dl\ |F_s) - P(\delta_i = 1,\, Z_i \in dl,\, T_n \leq t\ |F_s)$$

$$= \frac{P(\delta_i = 1,\, Z_i \in dl)}{P(Z_i \geq s)} - \frac{P(\delta_i = 1,\, Z_i \in dl)P^k(Z_1 \in [s,\, t])}{P^{k+1}(Z_1 \geq s)}$$

$$= \frac{P^k(Z_i \geq s) - P^k(Z_i \in [s,\, t])}{P^{k+1}(Z_i \geq s)}\ P(\delta_i = 1,\, Z_i \in dl)\ . \qquad (2.4)$$

On the other hand, we have

$$E[I_{[T_n > t]}(\Lambda(Z_i \wedge s+h) - \Lambda(Z_i \wedge s))|F_s] = \int_{(s,\infty]} [\Lambda(l \wedge s+h) - \Lambda(s)]P(Z_i \in dl,\, T_n > t \mid F_s) \quad ,$$

and

$$P(Z_i \in dl,\, T_n > t \,|F_s) = \begin{cases} \dfrac{P(Z_i \in dl)}{P(Z_i \geq s)} & \text{if} \quad l > t \\ \dfrac{P^k(Z_i \geq s) - P^k(Z_i \in [s,\, t])}{P^{k+1}(Z_i \geq s)} P(Z_i \in dl) & \text{if} \quad s < l \leq t \end{cases}$$

in a manner similar to the way (2.4) was derived.

If, however, we use $\dfrac{P^k(Z_i \geq s) - P^k(Z_i \in [s,\, t])}{P^{k+1}(Z_i \geq s)} P(Z_i \in dl)$ throughout the above expression (i.e., for both $s < l \leq t$ and $l > t$), we get a lower bound of the conditional expectation, namely

$$E\big[I_{[T_n > t\,]}[\Lambda(Z_i \wedge s{+}h) - \Lambda(Z_i \wedge s)]|F_s\big]$$

$$\geq \int_{(s,\,\infty]} [\Lambda(l \wedge s{+}h) - \Lambda(s)] \frac{P^k(Z_i \geq s) - P^k(Z_i \in [s,\, t])}{P^{k+1}(Z_i \geq s)}\, P(Z_i \in dl)$$

$$= \frac{P^k(Z_i \geq s) - P^k(Z_i \in [s,\, t])}{P^{k+1}(Z_i \geq s)} \int_{(s,\,\infty]} [\Lambda(l \wedge s{+}h) - \Lambda(s)] P(Z_i \in dl).$$

Now substituting this lower bound and (2.4) into (2.3), we have

$$E[I_{[T_n > t]}\, M_i(s + h) - M_i(s) \mid F_s] =$$

$$= \int_{(s,\, s+h]} P(\delta_i = 1,\, T_n > t,\, Z_i \in dl \mid F_s) - \int_{(s,\,\infty]} [\Lambda(l \wedge s{+}h) - \Lambda(s)] P(Z_i \in dl,\, T_n > t \mid F_s)$$

$$\leq \int_{(s,\, s+h]} \frac{P^k(Z_i \geq s) - P^k(Z_i \in [s,t])}{P^{k+1}(Z_i \geq s)}\, P(\delta_i = 1,\, Z_i \in dl) -$$

$$- \frac{P^k(Z_i \geq s) - P^k(Z_i \in [s,\, t])}{P^{k+1}(Z_i \geq s)} \int_{(s,\,\infty]} [\Lambda(l \wedge s + h) - \Lambda(s)] P(Z_i \in dl)$$

$$= \frac{P^k(Z_i \geq s) - P^k(Z_i \in [s,\, t])}{P^k(Z_i \geq s)} \left[\int_{(s,\, s+h]} \frac{P(\delta_i = 1,\, Z_i \in dl)}{P(Z_i \geq s)} - \int_{(s,\,\infty]} \frac{[\Lambda(l \wedge s{+}h) - \Lambda(s)] P(Z_i \in dl)}{P(Z_i \geq s)} \right]$$

$$= 0 \qquad a.s. \quad ,$$

where the last equality follows by the fact that the term in brackets is the conditional expectation of the martingale difference $E[M_i(s+h) - M_i(s)\,|F_s]$. ■

Lemma 2.2 Assume *X(t)* is nonnegative and predictable with respect to F_t , and suppose $\int_0^t X(s)dM(s)$ is well defined (i.e., $E\int_0^t X^2 d{<}M{>} < \infty$).

Then

$$EI_{[T_n > t]}\int_0^t X(s)dM(s) \le 0 \,.$$

Proof: Since $M = \sum_1^n M_i$, it follows that

$$E\, I_{[T_n > t]} \int_0^t X(s)dM(s) = \sum_1^n E\int_0^t I_{[T_n > t]}X(s)dM_i(s) \,.$$

Thus we need only show that for any i

$$E\int_0^t I_{[T_n > t]}X(s)dM_i(s) \;\le\; 0 \,.$$

To this end, let us write $\int_0^t I_{[T_n > t]}X(s)dM_i(s)$ as an L_2-limit of the summation $\sum_{t_j} X(t_{j-1})\, I_{[T_n > t]}\, \{M_i(t_j) - M_i(t_{j-1})\}$, and notice that by taking double expectation

$$E\left[\sum_{t_j} X(t_{j-1})\, I_{[T_n > t]}\{M_i(t_j) - M_i(t_{j-1})\}\right]$$

$$= E\left[\sum_{t_j} X(t_{j-1})E[I_{[T_n > t]}\{M_i(t_j) - M_i(t_{j-1})\}\,|F_{t_{j-1}}]\right] \le 0 \;,$$

where the last inequality follows by Lemma 2.1 and the assumption $X \ge 0$.

Thus the L_2-limit of the summation, $\int_0^t I_{[T_n > t]}\, X(s)dM(s)$, also has nonpositive mean. ■

Lemma 2.3

$$E(1 - \hat{F}_n(t))I_{[T_n > t]} \;\ge\; [1 - F(t)]EI_{[T_n > t]}$$

or equivalently,

$$E\frac{1-\hat{F}_n(t)}{1-F(t)}\, I_{[T_n > t]} \ge E I_{[T_n > t]}$$

or equivalently,

$$E[\hat{F}_n(t) - F(t)]\, I_{[T_n > t]} \le 0 \ \ .$$

Proof: By the representation (3.2.15) of Gill (1980), we have

$$\hat{F}_n(t) - F(t) = [1 - F(t)] \int_0^t \frac{1-\hat{F}_n(s-)}{1-F(s)} \frac{J}{Y} dM(s) \ - \ I_{[T_n < t]} \ \frac{[1-\hat{F}_n(T_n)][F(t) - F(T_n)]}{1-F(T_n)} \ .$$

Multiplying $I_{[T_n > t]}$ on both sides, we get

$$[\hat{F}_n(t) - F(t)] I_{[T_n > t]} = [1 - F(t)]\, I_{[T_n > t]} \int_0^t \frac{1-\hat{F}_n(s-)}{1-F(s)} \frac{J(s)}{Y(s)} dM(s) \ \ .$$

Now taking expectation on both sides and applying Lemma 2.2, we get that

$$E[\hat{F}_n(t) - F(t)]\, I_{[T_n > t]} = [1 - F(t)]E\, I_{[T_n > t]} \int_0^t \frac{1-\hat{F}_n(s-)}{1-F(s)} \frac{J(s)}{Y(s)} dM(s) \le 0 \ . \qquad \blacksquare$$

Lemma 2.4

$$E\,\frac{1-\hat{F}_n(T_n)}{1-F(T_n)} I_{[T_n \le t]} \le E I_{[T_n \le t]} = P(T_n \le t) \ \ .$$

Proof: It is well known that $\frac{1-\hat{F}_n(T_n \wedge t)}{1-F(T_n \wedge t)}$ is a martingale in t (see, e.g., Gill p.40 (1980)), thus we have

$$E\frac{1-\hat{F}_n(T_n \wedge t)}{1-F(T_n \wedge t)} = E\frac{1-\hat{F}_n(0)}{1-F(0)} = 1 \ .$$

Consequently,

$$E\frac{1-\hat{F}_n(T_n \wedge t)}{1-F(T_n \wedge t)}(I_{[T_n > t]} + I_{[T_n \le t]}) = 1 = E(I_{[T_n > t]} + I_{[T_n \le t]})$$

$$E\frac{1-\hat{F}_n(T_n \wedge t)}{1-F(T_n \wedge t)} I_{[T_n > t]} + E\frac{1-\hat{F}_n(T_n \wedge t)}{1-F(T_n \wedge t)} I_{[T_n \le t]} = E I_{[T_n > t]} + E I_{[T_n \le t]}$$

$$E\frac{1-\hat{F}_n(t)}{1-F(t)} I_{[T_n > t]} + E\frac{1-\hat{F}_n(T_n)}{1-F(T_n)} I_{[T_n \le t]} = E I_{[T_n > t]} + E I_{[T_n \le t]}$$

By Lemma 2.3 ,

$$E\frac{1-\hat{F}_n(t)}{1-F(t)}I_{[T_n>t]} \geq EI_{[T_n>t]} \;.$$

Thus it follows that

$$E\frac{1-\hat{F}_n(T_n)}{1-F(T_n)}I_{[T_n\leq t]} \leq EI_{[T_n\leq t]} = P(T_n \leq t) \;. \qquad \blacksquare$$

Now we are ready to prove Theorem 2.1.

PROOF OF THEOREM 2.1 :

We start with an expression of Gill (1980) [p. 38, (3.2.16)]

$$E[F(t)-\hat{F}_n(t)] = E\left[I_{[T_n<t]}\frac{[1-\hat{F}_n(T_n)]\;[F(t)-F(T_n)]}{1-F(T_n)}\right]$$

for t such that $F(t) < 1$.

From this, it is easy to calculate

$$E\{[F(t+dt)-\hat{F}_n(t+dt)]-[F(t)-\hat{F}_n(t)]\} \;. \qquad (2.5)$$

To do the calculation, we note that

$$E\{[F(t+dt)-\hat{F}_n(t+dt)]-[F(t)-\hat{F}_n(t)]\} = E\{F(t+dt)-F(t)-[\hat{F}_n(t+dt)-\hat{F}_n(t)]\}$$

$$= E\left[I_{[T_n<t+dt]}\frac{[1-\hat{F}_n(T_n)][F(t+dt)-F(T_n)]}{1-F(T_n)} - I_{[T_n<t]}\frac{[1-\hat{F}_n(T_n)][F(t)-F(T_n)]}{1-F(T_n)}\right]$$

and by writing the indicator $I_{[T_n<t+dt]}$ as $I_{[T_n<t]}+I_{[t\leq T_n<t+dt]}$, we get

$$E\left[I_{[T_n<t]}\frac{1-\hat{F}_n(T_n)}{1-F(T_n)}[F(t+dt)-F(t)] + I_{[t\leq T_n<t+dt]}\frac{1-\hat{F}_n(T_n)}{1-F(T_n)}[F(t+dt)-F(T_n)]\right]$$

It is easy to see that the above two terms are both ≥ 0 , hence the expectation is also ≥ 0.

On the other hand, if we change $[F(t+dt)-F(T_n)]$ to $[F(t+dt)-F(t)]$ in the second term above, we enlarge it ($t \leq T_n < t+dt$ therein), and thus get an upper bound given by

$$E\left[I_{[T_n<t]}\frac{1-\hat{F}_n(T_n)}{1-F(T_n)}[F(t+dt)-F(t)] + I_{[t\leq T_n<t+dt]}\frac{1-\hat{F}_n(T_n)}{1-F(T_n)}\;[F(t+dt)-F(t)]\right]$$

$$= E\left[I_{[T_n < t + dt]}\frac{1 - \hat{F}_n(T_n)}{1 - F(T_n)}\right][F(t + dt) - F(t)] \le P(T_n \le t + dt)[F(t + dt) - F(t)] \quad ,$$

where the last inequality follows from Lemma 2.4.

So we have

$$0 \le (2.5) \le P(T_n \le t + dt)[F(t + dt) - F(t)].$$

Now because the middle term of (2.1) (without taking expectation) is bounded above by

$$\sum_{t_j}\theta_*(t_j)\{[F(t_j + dt) - \hat{F}_n(t_j + dt)] - [F(t_j) - \hat{F}_n(t_j)]\} + \sum_{t_j}[\theta^*(t_j) - \theta_*(t_j)]\ [F(t_j + dt) - F(t_j)] \quad ,$$

where $\theta^*(t_j)$ and $\theta_*(t_j)$ are the maximum and minimum of θ over the small interval $[t_j, t_j + dt)$ respectively, it is easy to see that the expectation of this summation is bounded above by

$$\sum_{t_j}\theta_*(t_j)P(T_n \le t + dt)\ [F(t_j + dt) - F(t_j)] + \sum_{t_j}[\theta^*(t_j) - \theta_*(t_j)]\ [F(t_j + dt) - F(t_j)] \quad .$$

Upon taking limits, we see that the second summation above tends to zero in view of our assumption $\int I_{[\ \ ]}dF = 0$. The first summation above is easily seen to be bounded by $\int_0^\infty \theta(t)P^n(Z_1 \le t + \varepsilon)dF(t)$ for any $\varepsilon > 0$. Letting $\varepsilon \to 0$, we get the upper bound of (2.1). The lower bound 0 can be proved in a similar way and is easier. ■

Remark 2.2: From the above proof, we see that if we do not assume $\int_0^\infty I_{[\theta(t) \text{ is a discontinuity point }]}dF(t) = 0$, then the bounds in (2.1) are off at most by a ~~factor~~ term $\int \theta^+ - \theta_-\, dF$. Where $\theta^+(t) = \limsup_{s \to t}\ \theta(s)$, and $\theta_-(t) = \liminf_{s \to t}\ \theta(s)$.

The consistency of the statistic $\int_0^T [1 - \hat{F}_n(t)]dt$ is known (Susarla and Van Ryzin, 1979), (Gill, 1980). The consistency of the statistic of the form $\int_0^T \frac{w(t)}{1 - \hat{F}_n(t)}dt$ is the content of the next theorem. First we need the following lemma as conjectured by Gill (1980, p.40).

Lemma 2.6 For any $\beta \in (0, 1)$

$$P\{(1 - \hat{F}_n(t)) \geq \beta^2(1 - F(t));\ t \in [0, T)\} \geq 1 - \beta - \frac{e}{\beta} e^{-\frac{1}{\beta}}$$

Proof: Notice the fact that if we denote the Kaplan-Meier estimator of the censoring distribution $G(t)$ by $\hat{G}_n(t)$, then $[1 - \hat{F}_n(t)][1 - \hat{G}_n(t)]$ is just the usual empirical survival function corresponding to $[1 - F(t)][1 - G(t)]$.

Now since

$$P\{(1 - \hat{G}_n) \leq \frac{1}{\beta}(1 - G);\ t \in [0, T)\ \} \geq 1 - \beta$$

and

$$P\{(1 - \hat{F}_n)(1 - \hat{G}_n) \geq \beta(1 - F)(1 - G);\ t \in [0, T)\ \} \geq 1 - \frac{e}{\beta} e^{-\frac{1}{\beta}}$$

for $\beta \in (0, 1)$ (see Gill (1980), Wellner (1978), Gill (1983)), and

$$\{(1 - \hat{G}_n) \leq \frac{1}{\beta}(1-G);\ t \in [0, T)\} \cap \{(1 - \hat{F}_n)(1 - \hat{G}_n) \geq \beta\,(1 - F)(1 - G);\ t \in [0, T)\}$$

implies

$$\{(1 - \hat{F}_n) \geq \beta^2\,(1 - F);\ t \in [0, T)\};$$

the conclusion follows from Bonferroni's inequality. ■

Theorem 2.3 Suppose $w(t)$ is a real function and $\alpha > 0$, then

$$\int_0^T \frac{w(t)}{[1 - \hat{F}_n(t)]^\alpha} dt \overset{P}{\rightarrow} \int_0^\infty \frac{w(t)}{[1 - F(t)]^\alpha} dt \qquad as\ n \rightarrow \infty.$$

What about CLT

$\sqrt{n}\, [\int (\frac{w(t)}{1-\hat{F}} - \frac{w(t)}{1-F}) dt$

provided the right hand side is well defined.

Proof: By considering separately $w^+(t)$ and $w^-(t)$, we can and will assume, without loss of generality, that $w(t) \geq 0$. Let us first consider the case $\int_0^\infty \frac{w(t)}{[1 - F(t)]^\alpha} dt < \infty$.

In this case, the convergence of $\int_0^{M\wedge T} \frac{w(t)}{[1-\hat{F}_n(t)]^\alpha}dt \xrightarrow{P} \int_0^M \frac{w(t)}{[1-F(t)]^\alpha}dt$, where $1-F(M)>0$, is obvious. And by Lemma 2.6 we have, with probability no less than $1-\beta-\frac{e}{\beta}e^{-\frac{1}{\beta}}$, that

$$\frac{1}{\beta^2}\int_M^\infty \frac{w(t)}{[1-F(t)]^\alpha}dt \geq \int_{M\wedge T}^{T} \frac{w(t)}{[1-\hat{F}_n(t)]^\alpha}dt \ . \tag{2.6}$$

Now, for any given $\eta > 0$,

$$P\{\ |\int_0^T \frac{w(t)}{[1-\hat{F}_n(t)]^\alpha}dt - \int_0^\infty \frac{w(t)}{[1-F(t)]^\alpha}dt\ | \geq \eta\ \} \tag{2.7}$$

$$\leq P\{\ |\ \int_0^{T\wedge M} \frac{w(t)}{[1-\hat{F}_n(t)]^\alpha}dt - \int_0^M \frac{w(t)}{[1-F(t)]^\alpha}dt\ | \geq \frac{\eta}{3}\} +$$

$$+ P\{\ |\ \int_{T\wedge M}^{T} \frac{w(t)}{[1-\hat{F}_n(t)]^\alpha}dt\ | \geq \frac{\eta}{3}\} + P\{\ |\int_M^\infty \frac{w(t)}{[1-F(t)]^\alpha}dt\ | \geq \frac{\eta}{3}\}\ .$$

Choose β small enough, such that $1-\beta-\frac{e}{\beta}e^{-\frac{1}{\beta}} \geq 1-\frac{\eta}{2}$, and choose $M = M_\beta > 0$, such that $1-F(M)>0$, and $\frac{1}{\beta^2}\int_M^\infty \frac{w(t)}{[1-F(t)]^\alpha}dt \leq \frac{\eta}{3}$. Then the last term of (2.7) vanishes, the middle term is less than $\frac{\eta}{2}$ in view of (2.6) and the way we choose β and M_β. The last term of (2.7) approaches zero when $n \to \infty$, so we can choose $N > 0$ such that whenever $n > N$, this term is less then $\frac{\eta}{2}$. Thus the sum of the three probability terms in (2.7) is less than η. This completes the proof because η is arbitrary.

If $\int_0^\infty \frac{w(t)}{[1-F(t)]^\alpha}dt = +\infty$, the same truncation technique as in the proof of Theorem 2.2 applies, except here the truncation should be

$$w^K = min[\ w(t),\ K\ ]\ I_{[t<K]} \qquad \text{if } \tau_F = \infty$$

$$w^K = min[\ w(t),\ K\ ]\ I_{[t < \tau_F - \frac{1}{K}]} \qquad \text{if } \tau_F < \infty\ . \qquad \blacksquare$$

For the remainder of the thesis, we assume that our lifetime distribution $F(t)$ is continuous.

The next theorem is essentially the Theorem 2.1 of Gill (1983), with a slight generalization.

Theorem 2.4 Let $h(t)$ be a continuous function which admits a representation $h(t) = h_1(t) - h_2(t)$, where both $h_1(t)$ and $h_2(t)$ satisfy

$$(i)\ \ h_i \geq 0$$

$$(ii)\ \ h_i \textit{ nonincreasing on } [M,\tau_F] \textit{ for some } M < \tau_F \tag{2.8}$$

$$(iii)\ \ \int_0^{\tau_F} h_i^2(t)\ \frac{dF}{(1-F)^2(1-G)} < \infty\ .$$

Then the processes $(hZ)^T$, $(\int hdZ)^T$ and $(\int Zdh)^T$ converge jointly in $D[0,\tau_F]$ in distribution to processes $hZ^{(\infty)}$, $\int hdZ^{(\infty)}$ and $\int Z^{(\infty)}dh$ respectively, and

$$hZ^{(\infty)} = \int hdZ^{(\infty)} + \int Z^{(\infty)}dh \ \ ; \qquad h(t)Z^{(\infty)}(t) \overset{a.s.}{\to} 0 \quad as\ t \to \tau_F$$

where

$$Z = Z_n = \sqrt{n}\frac{1-\hat{F}}{1-F}\ ; \quad \text{and} \quad (\ *(t)\ )^T \textit{ means } (\ *(t \wedge T))$$

$$\text{and } Z^{(\infty)} = \textit{Brownian motion with clock } \ C(t) = \int_0^t \frac{dF}{(1-F)^2(1-G)}\ .$$

Proof: First, we show that the three limiting processes are well defined on $[0,\tau_F\ ]$.

Notice that $h = h_1 - h_2$, so the limiting processes

$$hZ^{(\infty)} = h_1Z^{(\infty)} - h_2Z^{(\infty)}$$

$$\int hdZ^{(\infty)} = \int h_1dZ^{(\infty)} - \int h_2dZ^{(\infty)}$$

$$\int Z^{(\infty)}dh = \int Z^{(\infty)}dh_1 - \int Z^{(\infty)}dh_2 \quad .$$

The right hand side is well defined by our assumption (2.8) and Gill (1983, Remark 2.2), hence our limiting processes are well defined.

The only thing remaining to be proved is the 'tightness at τ_F' , since the convergence on $[0, \tau_F - \varepsilon\,]$ is apparent, and the limiting processes do exist on $[0, \tau_F\,]$ as we have just shown. (see Billingsley, 1968)

The following facts are useful :

$$h^2 = (h_1 - h_2)^2 \le (h_1 + h_2)^2 \le 2h_1^2 + 2h_2^2 \quad ;$$

If $h^+ = h_1{+}h_2$, then h^+ satisfies (2.8) also, i.e.

(*i*) $h^+ \ge 0$

(*ii*) h^+ *nonincreasing on* $[M, \tau_F]$

(*iii*) $\int_0^{\tau_F} (h^+)^2 dC(t) < \infty$

We first show tightness of $\int hdZ$:

By Lenglart's inequality (see, e.g., Gill 1980, p.18 Th.2.4.2)

$$P\{\sup_{v \le t} |\int_v^t h(s)dZ(s)| > \varepsilon\ \} \ \le\ \frac{\eta}{\varepsilon^2} + P\{\int_v^t \frac{h^2(s)[1 - \hat{F}_n(s-)]^2}{[1 - F(s)]^2}\ \frac{n}{Y(s)} d\Lambda(s) > \eta\}$$

$$\le \frac{\eta}{\varepsilon^2} + P\{\int_v^t \frac{[h^+(s)]^2[1 - \hat{F}_n(s-)]^2}{[1 - F(s)]^2}\ \frac{n}{Y(s)} d\Lambda(s) > \eta\}$$

where the last inequality follows from the fact $[h^+(s)]^2 \ge [h(s)]^2$.

Now, the exact same argument as in Gill (1983, p.55, last eight lines), gives us the desired tightness.

Finally, to show the tightness of hZ, note that

$$\sup_{\tau \le t} |h(t)Z(t) - h(\tau)Z(\tau)| \tag{2.9}$$

$$\le \sup_{\tau \le t} |h(t)[Z(t) - Z(\tau)]| + \sup_{\tau \le t} |[h(t) - h(\tau)]Z(\tau)|$$

For the second term, we observe that since $h_i(t)$ is nonincreasing, we have $h_i(t) - h_i(\tau) \le 0$. Thus

$$\sup_{\tau \le t} |[h(t) - h(\tau)]Z(\tau)| =$$

$$= \sup_{\tau \le t} |[(h_1(t) - h_1(\tau)) - (h_2(t) - h_2(\tau))]Z(\tau)|$$

$$\le \sup_{\tau \le t} |(h_1(t) - h_1(\tau)) + (h_2(t) - h_2(\tau))||Z(\tau)|$$

$$= \sup |h^+(t) - h^+(\tau)||Z(\tau)|$$

$$\le (h^+(\tau) - h^+(\tau))|Z(\tau)|$$

where the last inequality follows by h^+ being nonincreasing.

For the first term,

$$\sup \left|h(t)[Z(t) - Z(\tau)]\right| \le \sup \left|h^+(t)[Z(t) - Z(\tau)]\right|$$

since $|h(t)| \le |h^+(t)|$.

Therefore equation (2.9) above is bounded by

$$\sup\left|h^+(t)[Z(t) - Z(\tau)]\right| \;+\; (h^+(t) - h^+(\tau))Z(\tau) \quad .$$

Now follow exactly the same argument (with h replace by h^+) as in Gill (1983, p.55, line 14) to give us the tightness. ■

Theorem 2.5

$$\sqrt{n} \quad [\int_0^{T_n} \theta(x)d\hat{F}_n(x) - \int_o^{T_n} \theta(x)dF(x)] \xrightarrow{D} N(0, \sigma^2)$$

as $n \to \infty$, provided

(*i*) $\sigma^2 = \int_0^\infty \{\theta(x)[1-F(x)] - \overbrace{\int_x^\infty \theta(t)dF(t)}\}^2 \frac{dF(x)}{[1-F(x)]^2[1-G(x)]} < \infty$

$$\int_x^\infty \theta(t)d(1-F(t))$$
$$= \theta(t)[1-F(t)]\Big|_x^\infty - \int_x^\infty 1-F(t)d\theta(t)$$
$$= -\theta(x)[1-F(x)] - \int_x^\infty 1-F(t)d\theta(t)$$

(*ii*) $\theta(t)[1 - F(t)] - \int_t^\infty \theta(s)dF(s)$ *is well defined* and *regular at* τ_F

(*i.e., it admits a decomposition* $h_1 - h_2$ *where both* h_1 and h_2 *satisfy* (2.8).)

So, (*i*) can also be written as

$$\sigma^2 = \int_0^\infty \left\{ \int_x^\infty [1-F(t)]d\theta(t) \right\}^2 \frac{dF}{[1-F(x)]^2[1-G(x)]}$$

$$(iii)\quad [\int_t^{\infty}\theta dF]^2 \int_0^t \frac{dF}{(1-F)^2(1-G)} \to 0 \quad as \ t \to \infty$$

$$(iv)\quad \int_0^{\infty}\theta(t)dF(t) < \infty \quad .$$

Proof: Observe that

$$d\frac{\hat{F}_n(x)-F(x)}{1-F(x)}(1-F(x)) = 1-F(x)d\frac{\hat{F}_n(x)-F(x)}{1-F(x)} + \frac{\hat{F}_n(x)-F(x)}{1-F(x)}d(1-F(x)) \quad .$$

So we have

$$\int_0^{T_n}\theta(x)d\hat{F}_n(x) - \int_0^{T_n}\theta(x)dF(x) = \int_0^{T_n}\theta(x)d(\hat{F}_n(x)-F(x))$$

$$= \int_0^{T_n}\theta(x)d\{\frac{\hat{F}_n(x)-F(x)}{1-F(x)}(1-F(x))\}$$

$$= \int_0^{T_n}\theta(x)[1-F(x)]d\frac{\hat{F}_n(x)-F(x)}{1-F(x)} + \int_0^{T_n}\frac{\hat{F}_n(x)-F(x)}{1-F(x)}\theta(x)d(1-F(x)) \quad . \tag{2.10}$$

Define

$$h(t) = \int_t^{\infty}\theta(x)d(1-F(x)\ )$$

(which is well defined by the assumption *(ii)*). Integration by parts in the second term in (2.10) allows us to rewrite (2.10) as

$$= \int_0^{T_n}\theta(x)[1-F(x)]d\frac{\hat{F}_n(x)-F(x)}{1-F(x)} + \frac{\hat{F}_n(x)-F(x)}{1-F(x)}h(x)\Big|_0^{T_n} - \int_0^{T_n}h(x)d\frac{\hat{F}_n(x)-F(x)}{1-F(x)}$$

$$= \int_0^{T_n}\{\theta(x)[1-F(x)] - h(x)\}d\frac{\hat{F}_n(x)-F(x)}{1-F(x)} + \frac{\hat{F}_n(T_n)-F(T_n)}{1-F(T_n)}h(T_n)$$

Now Theorem 2.4 and assumption *(iii)* implies

$$\sqrt{n}\ \frac{\hat{F}_n(T_n)-F(T_n)}{1-F(T_n)}h(T_n) \overset{P}{\to} 0\ , \qquad \text{and}$$

$$\sqrt{n}\ \int_0^{T_n}\{\ \theta(x)[1-F(x)] - h(x)\ \}d\frac{\hat{F}_n(x)-F(x)}{1-F(x)} \overset{D}{\to} N(0,\sigma^2) \quad .$$

Therefore the conclusion of this theorem is true. ■

CHAPTER 7

DIFFERENCE OF MEANS TEST

Despite the development of many types of nonparametric rank tests for the two sample problem in the no censoring case, the t-test still retains its primer position as the standard two sample test.

We will formulate here an analog of the t-test in the censoring case and state a few results about its asymptotic properties. Let us agree to call it the 'difference of means test'.

The test statistic is obviously to be given by (in the notation defined in Chapter three),

$$M_N = \sqrt{N}\,[\int_0^{T_1} (1 - \hat{F}_1)dt - \int_0^{T_2} (1 - \hat{F}_2)dt] \quad . \tag{7.1}$$

Because of the work of Susarla and Van Ryzin (1980) and Gill (1983) on the mean survival time estimator, the asymptotic theory for the test statistic M_N is readily formulated.

In another formulation of the problem, we can use the results on regression analysis (with censored data) to yield different kinds of two sample difference of means tests. We will compare these methods and study their relationships.

Theorem 7.1 Using all the notation of Chapter three for two samples, we have

$$\sqrt{N}\left[\int_0^{T_1}(1-\hat{F}_1)dt - \int_0^{T_1}(1-F_1)dt - \int_0^{T_2}(1-\hat{F}_2)dt + \int_0^{T_2}(1-F_2)dt\right] \overset{D}{\rightarrow} N(0,\, \sigma_A^2) \quad , \tag{7.2}$$

whenever its asymptotic variance σ_A^2 is finite, where

$$\sigma_A^2 = \frac{1}{\lambda_1}\int_0^\infty [\int_t^\infty 1 - F_1 ds]^2 \frac{dF_1(t)}{(1-F_1)^2(1-G_1)} + \frac{1}{\lambda_2}\int_0^\infty [\int_t^\infty 1 - F_2 ds]^2 \frac{dF_2(t)}{(1-F_2)^2(1-G_2)} \quad .$$

Here as before λ_i denotes the limit of $\frac{n_i}{N}$. We sometimes use the equivalent notation $\lambda = lim \frac{n_1}{N}$ and $1 - \lambda = lim \frac{n_2}{N}$.

Proof : Because the two samples are independent, we can apply Gill (1983)[Theorem 2.1] to each sample and make use of the independence to finish the proof. ■

Koul, Susarla and Van Ryzin (1981) suggest in the context of linear regression that one treat

$$Y_{ij}^* = \frac{\delta_{ij} Z_{ij}}{1 - \hat{G}_i(Z_{ij})} \qquad i = 1, 2;\ j = 1, \cdots, n_i \tag{7.3}$$

as the observation and apply the usual least square regression procedure to these data. They proved the consistency and asymptotic normality of this method. Applying this method to our case yields the two sample test statistic

$$\frac{1}{n_1}\sum_j Y_{1j}^* - \frac{1}{n_2}\sum_j Y_{2j}^* \quad . \tag{7.4}$$

Leurgans (1984) and Zheng (1984) suggest that instead of (7.3), we use

$$Y_{ij}^* = \int_0^{Z_{ij}} \frac{dt}{1 - \hat{G}_i(t)} = \int_0^\infty \frac{I_{[Z_{ij} > t]}}{1 - \hat{G}_i(t)} dt \quad (set\ \ \frac{0}{0} = 0) \ , \tag{7.5}$$

$$i = 1,2;\ j = 1, \cdots, n_i$$

as the observations, called 'synthetic data' by Leurgans (1984). Both suggest that (7.5) is better then (7.3).

However, in the two sample situation, we find that these two methods are exactly the same as shown below.

Theorem 7.2 Both (7.3) and (7.5) are exactly the same test statistics as M_N in Theorem 7.1. (see equation (7.1)) provided we always treat the last observation as uncensored in the Koul, Susarla and Van Ryzin (1981) procedure.

Proof: If we write out the resulting statistic by applying Leurgans' synthetic data, we see it is given by

$$\sqrt{N}\left[\frac{1}{n_1}\sum_j\int_0^{Z_{1j}}\frac{dt}{1-\hat{G}_1(t)}-\frac{1}{n_2}\sum_j\int_0^{Z_{2j}}\frac{dt}{1-\hat{G}_2(t)}\right]$$

$$=\sqrt{N}\left[\frac{1}{n_1}\sum_j\int_0^{\infty}\frac{I_{[Z_{1j}>t]}}{1-\hat{G}_1(t)}dt-\frac{1}{n_2}\sum_j\int_0^{\infty}\frac{I_{[Z_{2j}>t]}}{1-\hat{G}_2(t)}dt\right]$$

$$=\sqrt{N}\left[\int_0^{\infty}\frac{1}{n_1}\sum_j I_{[Z_{1j}>t]}\frac{dt}{1-\hat{G}_1(t)}-\int_0^{\infty}\frac{1}{n_2}\sum_j I_{[Z_{2j}>t]}\frac{dt}{1-\hat{G}_2(t)}\right]$$

$$=\sqrt{N}\left[\int_0^{\infty}(1-\hat{H}_1(t))\frac{dt}{1-\hat{G}_1(t)}-\int_0^{\infty}(1-\hat{H}_2(t))\frac{dt}{1-\hat{G}_2(t)}\right] \qquad (7.6)$$

Now apply the fact that $(1-\hat{F})(1-\hat{G})=(1-\hat{H})$, i.e., that the Kaplan-Meier estimator of the lifetime distribution multiplied by the Kaplan-Meier estimator of the censoring distribution is the usual empirical survival function estimator of $(1-H)=(1-F)(1-G)$; and remember for $t>T$ either $1-\hat{F}=0$ or $1-\hat{G}=0$ and $\frac{0}{0}=0$, we see that (7.6) becomes

$$\sqrt{N}\left[\int_0^{\infty}(1-\hat{F}_1)(1-\hat{G}_1)\frac{dt}{1-\hat{G}_1(t)}-\int_0^{\infty}(1-\hat{F}_2)(1-\hat{G}_2)\frac{dt}{1-\hat{G}_2(t)}\right]$$

$$=\sqrt{N}\left[\int_0^{T_1}(1-\hat{F}_1)dt-\int_0^{T_2}(1-\hat{F}_2)dt\right] \quad ,$$

which is exactly the same as M_N.

For the Koul, Susarla and Van Ryzin transformation Y^*, the test statistic becomes

$$\sqrt{N}\left[\frac{1}{n_1}\sum_j \frac{\delta_{1j}\, Z_{1j}}{1-\hat{G}_1(Z_{1j})} - \frac{1}{n_2}\sum_j \frac{\delta_{2j}\, Z_{2j}}{1-\hat{G}_2(Z_{2j})}\right] .$$

We first take a look at $\int_0^{T_i}(1-\hat{F}_i)dt$. Observe that

$$\int_0^{T_i}(1-\hat{F}_i)dt = \sum_j X_{ij}\, \Delta\hat{F}_i(X_{ij})$$

provided we always treat the last observation as an uncensored one (death).

Now note that the jump size of the Kaplan-Meier estimator at X_{ij} is $\frac{1}{n_i}\frac{1}{1-\hat{G}_i(X_{ij})} = \Delta\hat{F}_i(X_{ij})$ (see Susarla, Tsai and Van Ryzin (1984)), so that the above summation can be rewritten as

$$\sum X_{ij}\,\frac{1}{n_i}\frac{1}{1-\hat{G}_i(X_{ij})} = \sum \delta_{ij}\, Z_{ij}\,\frac{1}{n_i}\frac{1}{1-\hat{G}_i(X_{ij})} = \frac{1}{n_i}\sum\frac{\delta_{ij}\, Z_{ij}}{1-\hat{G}_i(Z_{ij})} .$$

The remainder of the proof is now clear. ■

If we believe, in some cases, that the censoring mechanisms are the same for the two samples, i.e., the censoring distributions are the same for both the drug and the placebo, we can pool the two samples to get a better estimator of the censoring distribution. One would naturally think that this will therefore give rise to a better test. We find, however, this is not the case.

Let us denote the pooled Kaplan-Meier estimator of the censoring distribution by $\overline{G}_N(t)$ to distinguishing it from $\hat{G}_i(t)$, the Kaplan-Meier estimator based only on sample i . Then the pooled version of the test statistic is (just replace $\hat{G}$ by $\overline{G}$)

$$\sqrt{N}\left\{\frac{1}{n_1}\sum_i\int_0^\infty \frac{I_{[Z_{1i}>t]}}{1-\overline{G}_N(t)}dt - \frac{1}{n_2}\sum_i\int_0^\infty \frac{I_{[Z_{2i}>t]}}{1-\overline{G}_N(t)}dt\right\} = \overline{M}_N \quad . \qquad (7.7)$$

The next theorem assures us that $\overline{M}_N$ is asymptotically normally distributed.

We will pay special attention to its variance.

Theorem 7.3 The statistic $\overline{M}_N$ as defined in (7.7) is asymptotically normally distributed (if properly normalized) with variance

$$\sigma^2 = \sigma_A^2 + \sigma_{B1}^2 + \sigma_{B2}^2 \quad ,$$

provided the following conditions are satisfied:

$$(i) \quad \sigma^2 < \infty \ ,$$

where σ_A^2 is defined as in Theorem 7.1, and ($i \neq i'$, $i, i' = 1, 2$.)

$$\sigma_{Bi}^2 = \int_0^\infty [\frac{\{(\int_t^\infty 1 - F_i ds) - (\int_t^\infty 1 - F_{i'} ds)\}}{\lambda_i(1 - H_i) + \lambda_{i'}(1 - H_{i'})} + (\int_t^\infty 1 - F_i ds)\frac{1}{1 - H_i}\ \frac{1}{\lambda_i}]^2 \lambda_i(1 - H_i) d\Lambda^C \quad ,$$

$$(ii) \quad \sqrt{N} \int_{T_i}^\infty (1 - F_i) ds \overset{P}{\rightarrow} 0 \ ; \quad as \ N \rightarrow \infty \ , i = 1, 2 \ ;$$

$$(iii) \quad \int_0^\infty [\int_t^\infty (1 - F_i) ds]^2 \frac{dG(t)}{(1 - G)^2 (1 - F^+)} < \infty \quad ,$$

where $F^+ = \lambda_1 F_1 + \lambda_2 F_2$.

In order to prove the theorem, we need the following results.

Theorem 7.4 The compensated counting processes (martingales) $M = N^D - \int Y d\Lambda^D$ and $L = N^C - \int Y d\Lambda^C$ are uncorrelated, i.e., $<M, L> = 0$; or, $M\,L$ is also a martingale.

Where

$$N^D(t) = \#\{ \ i : Z_i \leq t \ , \ \delta_i = 1 \ \} \ , \qquad \Lambda^D(t) = \int_{[0,\ t]} \frac{dF(s)}{1 - F(s-)}$$

$$N^C(t) = \#\{ \ i : Z_i \leq t \ , \ \delta_i = 0 \ \} \ , \qquad \Lambda^C(t) = \int_{[0,\ t]} \frac{dG(s)}{1 - G(s-)} \quad .$$

Proof: Notice that $M + L$ is also a compensated counting process, with intensity

$$\int Y d\Lambda^D + \int Y d\Lambda^C = \int Y d(\Lambda^D + \Lambda^C) \quad .$$

Thus, the following, denoted MA, is also a martingale,

$$MA = (M + L)^2 - \int Y d(\Lambda^D + \Lambda^C)$$

With a little bit of algebra, we see

$$M^2 + L^2 + 2ML - \int Y d\Lambda^D - \int Y d\Lambda^C = MA$$

$$\left[M^2 - \int Y d\Lambda^D\right] + \left[L^2 - \int Y d\Lambda^C\right] + 2ML = MA \quad .$$

Hence,

$$2ML = MA - \left[M^2 - \int Y d\Lambda^D\right] - \left[L^2 - \int Y d\Lambda^C\right]$$

is also a martingale. ■

Corollary The processes $\int_0 QdM$ and $\int_0 RdL$ are martingales with a zero correlation process, where we assume Q and R are predictable. Also, $\frac{\hat{F}_i - F_i}{1 - F_i}$ and $\frac{\hat{G}_i - G_i}{1 - G_i}$ are thus martingales with a zero correlation process. ■

Proof of Theorem 7.3:

We will use the notation M_i^D , M_i^C , M_1^+ , M_C^+ defined as follows:

$$Y^+ = Y_1 + Y_2 \quad ;$$

$$M_i^D = N_i^D - \int Y_i d\Lambda_i^D \quad ; \quad M_i^C = N_i^C - \int Y_i d\Lambda_i^C \quad ;$$

$$M_1^+ = [N_1^D + N_1^C] - \int Y_1 d[\Lambda_1^D + \Lambda^C\,] \quad ; \quad M_C^+ = [N_1^C + N_2^C] - \int Y^+ d\Lambda^C \quad ,$$

where $\Lambda^C = \int_{[0,\,t]} \frac{dG(s)}{1 - G(s-)}$. After a bit of algebra, one can easily prove that

$$M_1^+ = M_1^C + M_1^D \quad \text{and} \quad M_C^+ = M_1^C + M_2^C \quad .$$

Let us first look at

$$\frac{1}{n_1}\sum_i\int_0^\infty \frac{I_{[Z_{1i}>t]}}{1-\overline{G}_N(t)}dt - \int_0^{T_1}(1-F_1)dt \quad . \tag{7.8}$$

By a bit of algebra, we see that,

$$(7.8) = \int_0^\infty \frac{1}{n_1}\sum_i I_{[Z_{1i}>t]}\,\frac{dt}{1-\overline{G}_N(t)} - \int_0^{T_1}(1-F_1)dt$$

$$= \int_0^{T_1}\frac{1-\hat{H}_1}{1-\overline{G}_N}dt - \int_0^{T_1}(1-F_1)dt$$

$$= \int_0^{T_1}\frac{1-\hat{H}_1}{1-H_1}\,\frac{1-G}{1-\overline{G}_N}\,(1-F_1)dt - \int_0^{T_1}(1-F_1)dt$$

$$= \int_0^{T_1}\left[\frac{1-\hat{H}_1}{1-H_1}\,\frac{1-G}{1-\overline{G}_N} - 1\right](1-F_1)dt \quad .$$

By writing

$$\frac{1-\hat{H}_1}{1-H_1} = 1 + \frac{H_1-\hat{H}_1}{1-H_1} \quad , \quad \frac{1-G}{1-\overline{G}_N} = 1 + \frac{\overline{G}_N-G}{1-G} + \varepsilon^* \quad ,$$

we can write the equality as

$$(7.8) = \int_0^{T_1}\left[\frac{H_1-\hat{H}_1}{1-H_1} + \frac{\overline{G}_N-G}{1-G}\right](1-F_1)dt + \int_0^{T_1}\varepsilon^*(1-F_1)dt +$$

$$+ \int_0^{T_1}\frac{H_1-\hat{H}_1}{1-H_1}\,\frac{\overline{G}-G}{1-\overline{G}}(1-F_1)dt \quad .$$

Upon integration by parts, we get

$$(7.8) = \int_0^{T_1}[\int_t^\infty(1-F_1)ds]\;d\frac{H_1-\hat{H}_1}{1-H_1} + \int_0^{T_1}[\int_t^\infty(1-F_1)ds]\;d\,\frac{\overline{G}_N-G}{1-G}$$

$$+ \int_0^{T_1}\varepsilon^*(1-F_1)dt + \int_0^{T_1}\frac{H_1-\hat{H}_1}{1-H_1}\,\frac{\overline{G}_N-G}{1-\overline{G}_N}(1-F_1)dt$$

$$+ (\int_{T_1}^\infty(1-F_1)dt)\;(\frac{\hat{H}_1-H_1}{1-H_1} + \frac{G-\overline{G}_N}{1-G})_{\textit{evaluated at } T_1} \quad . \tag{7.9}$$

Now, the first two terms on the right hand side of the above can be written as

$$\int_0^{T_1} (\int_t^\infty 1 - F_1 ds)\frac{1}{1 - H_1}\ \frac{1}{n_1} dM_1^+(t) \ + \ \int_0^{T_1} (\int_t^\infty 1 - F_1 ds)\frac{1 - \overline{G}_-}{1 - G}\ \frac{1}{Y^+(t)} dM_c^+(t) \quad . \ (7.10)$$

Recall that

$$M_1^+ = M_1^C + M_1^D \qquad \text{and} \qquad M_C^+ = M_1^C + M_2^C \quad ,$$

and notice that by Theorem 7.4, $\{M_1^D, M_1^C, M_2^D, M_2^C\}$ are mutually uncorrelated multivariate counting processes. Hence, we finally have

$$(7.10) = \int_0^{T_1}(\int_t^\infty 1 - F_1 ds)\frac{1}{1 - H_1}\ \frac{1}{n_1} dM_1^D \ + \ \int_0^{T_1}(\int_t^\infty 1 - F_1 ds)\ \frac{1 - \overline{G}_-}{1 - G}\ \frac{1}{Y^+(t)} dM_2^C \ +$$

$$+ \ \int_0^{T_1}(\int_t^\infty 1 - F_1 ds)\left[\frac{1}{1 - H_1}\ \frac{1}{n_1} \ + \ \frac{1 - \overline{G}_-}{1 - G}\ \frac{1}{Y^+}\right] dM_1^C \quad .$$

Applying the same decomposition to the second sample term of the statistics, $\frac{1}{n_2}\sum_j \int_0^\infty \frac{I_{[Z_{2j} > t]}}{1 - \overline{G}_N(t)} dt - \int_0^{T_2} 1 - F_2 dt$, we will get the same thing except the subscript 1 is now 2.

And so we get

$$\sqrt{N}\ [\frac{1}{n_1}\sum_j \int_0^\infty \frac{I_{[Z_{1j} > t]}}{1 - \overline{G}_N} dt \ - \ \frac{1}{n_2}\sum_j \int_0^\infty \frac{I_{[Z_{2j} > t]}}{1 - \overline{G}_N} dt\ ]$$

$$= \sqrt{N}\Big[\ (A_1 - A_2) + (B_1 - B_2) + \ D \ \Big] \quad ,$$

where

$$A_i = \int_0^{T_i}(\int_t^\infty 1 - F_i ds)\frac{1}{1 - H_i}\frac{1}{n_i} dM_i^D$$

$$B_i = \int_0^{T_i}\left[\{\ (\int_t^\infty 1 - F_i ds) - I_{[t < T_{i'}]}\int_t^\infty 1 - F_{i'} ds\ \}\ \frac{1 - \overline{G}_-}{1 - G}\ \frac{1}{Y^+} + \ (\int_t^\infty 1 - F_i ds)\frac{1}{1 - H_i}\ \frac{1}{n_i}\ \right] dM_i^C$$

$$\text{for} \ \ i, i' = 1, 2 \qquad i \neq i'$$

$$D = \sum_{i=1,2}(-1)^{i+1}\int_0^{T_i}\varepsilon^*(1 - F_i)dt + \sum_{i=1,2}(-1)^{i+1}\ [\int_{T_i}^\infty(1 - F_i)dt](\frac{\hat{H}_i - H_i}{1 - H_i} + \frac{G - \overline{G}}{1 - G})_{\textit{evaluated at } T_i}$$

$$+ \sum_{i=1,2}(-1)^{i+1}\int_0^{T_1}\frac{H_1-\hat{H}_1}{1-H_1}\frac{\overline{G}_N-G}{1-\overline{G}_N}(1-F_1)dt \quad .$$

By the multivariate version of Rebolledo's theorem, (Anderson and Borgon (1985) or Gill (1980, Th.2.4.1) $\surd N[(A_1-A_2) + (B_1-B_2)]$ converges in distribution to a normal random variable with variance

$$\sigma^2 = \sigma_A^2 + \sigma_{B1}^2 + \sigma_{B2}^2 \quad .$$

Finally we have to show $\surd ND \overset{P}{\to} 0$.

The second sum of D is easy, since both

$$\frac{\hat{H}_i - H_i}{1-H_i} \quad \text{and} \quad \frac{G-\overline{G}}{1-G}$$

are bounded in probability (see Lemma 5.2) and we assumed $\surd N\int_{T_i}^{\infty} 1 - F_i ds \overset{P}{\to} 0$.

Thus, we see that

$$\surd N \sum_{i=1,2}(-1)^{i+1}\left[\int_{T_i}^{\infty} 1-F_i dt\right]\left(\frac{\hat{H}_i-H_i}{1-H_i}+\frac{G-\overline{G}}{1-G}\right)_{\textit{evaluated at } T_i} \overset{P}{\to} 0 \quad .$$

As for the first summation of D, the ε^* is nothing but

$$\varepsilon_i^* = \frac{\overline{G}-G}{1-G}\,\frac{\overline{G}-G}{1-\overline{G}} = \frac{(\overline{G}-G)^2}{(1-G)^2}\,\frac{1-G}{1-\overline{G}} \quad .$$

By Lemma 2.6 in Chapter two, we know $\frac{1-G}{1-\overline{G}}$ is bounded in probability, therefore,

$$P\{\surd N\int_0^{T_i}\varepsilon_i^*(1-F_i)ds > \eta\} = P\{\surd N\int_0^{T_i}\left(\frac{\overline{G}-G}{1-G}\right)^2\frac{1-G}{1-\overline{G}}(1-F_i)ds > \eta\}$$

$$\le \beta + \frac{e}{\beta}e^{-\frac{1}{\beta}} + P\{\int_0^{T_i}\frac{1}{\beta^2}\surd N\left(\frac{\overline{G}-G}{1-G}\right)^2(1-F_i)ds > \eta\}$$

$$\le \beta + \frac{e}{\beta}e^{-\frac{1}{\beta}} + P\{\int_0^{\tau^*}\frac{1}{\beta^2}\surd N\left(\frac{\overline{G}-G}{1-G}\right)^2(1-F_i)ds > \frac{\eta}{2}\} +$$

$$+ P\{\int_{\tau^*}^{T_i}\frac{1}{\beta^2}\surd N\left(\frac{\overline{G}-G}{1-G}\right)^2(1-F_i)ds > \frac{\eta}{2}\} \quad .$$

Again by Lemma 5.2, this is bounded by

$$\leq \beta + \frac{e}{\beta} e^{-\frac{1}{\beta}} + P\{\int_0^{\tau^*} \frac{1}{\beta^2} \sqrt{N} (\frac{\overline{G} - G}{1 - G})^2 (1 - F_i) ds > \frac{\eta}{2}\} +$$

$$+ \beta + P\{\int_{\tau^*}^{T_i} \frac{1}{\beta^2} \frac{1}{\beta} \sqrt{N} (\frac{\overline{G} - G}{1 - G})(1 - F_i) ds > \frac{\eta}{2}\} \quad . \qquad (7.11)$$

The last probability term in (7.11) can be made arbitrary small by choosing τ^* large and when N is large because of assumption *(iii)*. The other probability term of (7.11) apparently goes to zero because $\sqrt{N}(\frac{\overline{G} - G}{1 - G})^2 \xrightarrow{P} 0$ there. Finally, as the β'*s* are arbitrary numbers in $(0, 1)$ we can make the β terms arbitrary small. Therefore, the first summation of D is negligible.

For the last summation of D, we can use the fact that $\frac{1 - G}{1 - \overline{G}}$ is also bounded in probability (see Lemma 2.6 of Chapter two), and then almost the same argument as in the proof for the first summation also works for this last summation. ■

Remark 7.1: Pooled or not pooled? The theorems above say that in the two sample situation we should not pool to estimate the censoring distribution when using the 'synthetic data' method. If we do, the variance will be larger.

Remark 7.2: If we know the actual censoring distribution G (or G_1 and G_2 if they are not equal), and use it in the synthetic data expression, it is straightforward to find its limiting distribution by employing the similar representations as in the proof of Theorem 7.3. Surprisingly, this new statistic has a variance $\sigma^2 = \sigma_A^2 + \sigma_D^2$ which is bigger then σ_A^2 as in (7.2), which means *'we are better off not knowing G'* or *'although we know G , we'd rather use an estimator of it instead of the true G'*. See Koul, Susarla and Van Ryzin (1980 Remark 4.5) for a similar remark.

Remark 7.3: The method used here to prove Theorem 7.3 can also be used to investigate the 'synthetic data' method applied to the censoring regression problem.

CHAPTER 8

COMPARISON OF TESTS

In order to get some feeling about the small and moderate sample behavior of the tests proposed, we did some Monte Carlo simulations of the following situation:

sample size :	$n_1 = n_2 = 50$
null distribution :	$F_0 = 1 - e^{-t^2}$
censoring distribution :	$G_i = 1 - e^{-(\frac{t}{2})^2}$
alternatives :	$F_A = 1 - e^{-(\frac{t}{\lambda})^2}$

We also include the Mantel-Haenszel or log-rank test in our simulation comparison as a standard one. The 'mean' entry is the difference of means test as proposed in Chapter seven, the 'RANK' entry is the rank test defined by (3.3) as in Chapter three with the choice of the optimal J function given by Theorem 4.2. It is (4.7) with $\alpha = \frac{1}{4}$ in this particular setting. See table 1.

Except the first row, each value in table 1 is the average of 90,000 runs. The approximate five percent level for the first row there is set by adjusting the 1.96 significance level a little bit and running the simulation 160,000 times for each test. The random numbers are generated by calling the IMSL library subrouting GGWIB and using the same seed 123457.0 .

We sometimes have two values in one entry of the table. This is because we have different versions of the test. (see, e.g., Gill (1980, pp. 47-48) for two ways of estimating the variance). In the log-rank column, the unbracketed value is derived from the test using the first variance estimator of Gill (1980), while the bracket value is derived from the test using the second variance estimator there, which is the same one as suggested by Mantel (1966).

POWER SIMULATIONS FOR
PROPORTIONAL HAZARD ALTERNATIVES

	logrank	RANK	mean	censoring%
H_0: λ=1.0	5.043% (5.058%)	5.041% (5.041%)	5.039%	20.0%
H_A: λ=1.1	13.03% (13.05%)	13.39% (13.65%)	12.98%	23.2%
H_A: λ=1.2	34.28% (34.31%)	34.78% (35.02%)	34.42%	26.4%
H_A: λ=1.3	59.88% (59.98%)	60.04% (60.21%)	60.28%	29.7%
H_A: λ=1.4	80.27% (80.31%)	79.88% (80.01%)	80.74%	32.9%
H_A: λ=1.5	91.68% (91.76%)	91.20% (91.39%)	92.10%	36.0%

table 1

The RANK column also has double entries. The unbracketed value is the test defined in (3.3) and using the variance estimator of Chapter six. The bracketed value is the test statistic defined in Chapter three with the H_N now taken to be the pooled Kaplan-Meier estimator of the lifetime distribution.

The mean column is the difference of means test introduced in Chapter seven, with the variance estimator

$$\int_0^\infty [\int_t^{T_1} 1 - \hat{F}_1 ds]^2 \frac{dN_1(t)}{Y_1(t)[Y_1(t) - 1]} + \int_0^\infty [\int_t^{T_2} 1 - \hat{F}_2 ds]^2 \frac{dN_2(t)}{Y_2(t)[Y_2(t) - 1]} \quad . \tag{8.1}$$

The last column reports the percentage of censoring under the null and alternative hypotheses. We see from the table 1 that the three tests have almost identical power in this situation. Recall that in this situation the Mantel-Haenszel test is asymptotically most powerful.

POWER SIMULATIONS FOR

NON-PROPORTIONAL HAZARD ALTERNATIVES

	logrank	RANK	mean	censoring%
H_0: β = 2.0	5.043% (5.058%)	5.041% (5.041%)	5.039%	20.0%
H_A: β = 6.0	5.00% (6.60%)	4.82% (5.29%)	11.08%	19.7%
H_A: β = 4.0	5.09% (6.00%)	4.80% (5.12%)	6.89%	19.3%
H_A: β = 1.0	4.59% (5.08%)	4.87% (4.80%)	6.83%	24.2%
H_A: β = 0.8	4.36% (5.02%)	4.88% (4.76%)	8.96%	26.0%
H_A: β = 0.6	4.25% (5.09%)	4.82% (4.79%)	13.22%	28.3%
H_A: β = 0.4	4.43% (5.51%)	5.10% (5.27%)	19.94%	30.9%

table 2

Table 2 reports simulation results for non-proportional hazard alternatives; namely, we take the situation where every setting is the same as before except here we take

$$\text{alternatives :} \qquad F_A = 1 - e^{-(t)^{\beta}} \ ; \beta \neq 2 \ ,$$

that is, the Weibull distribution with different shape parameters.

This table shows that the difference of means test is better than either the Mantel-Haenszel or our rank test while the latter two are roughly the same. Both rank tests behave badly in this case because both of them took the wrong weight function.

The actual critical values used in the above two tables are not 1.96, and is reported in the following table. The actual level for using the 1.96 as a critical value is also reported there, where each entry is based on 160,000 runs.

significance level			
	logrank	RANK	mean
actual	1.928 (1.974)	1.860 (1.870)	2.037
level of 1.96	4.67% (5.21%)	3.87% (3.92%)	5.96%

table 3

Now, we present some theoretical comparison results.

The K-class of tests as defined in Gill (1980) can be made most powerful in the equal censoring situation by chosing the optimal weight (see Gill, 1980, p. 118). In the unequal censoring case we have the following lemma.

Lemma 8.1 For any test whose efficacy depends on the censoring distributions G_1 and G_2 only through the combination

$$\frac{(1 - G_1(t)\,)\,(1 - G_2(t)\,)}{(1 - \lambda)\,(1 - G_2(t)\,) + \lambda\,(1 - G_1(t)\,)} = 1 - G^*(t) \quad (say), \tag{8.2}$$

then its efficacy can not be better than the best one in the K-class.

Proof: Suppose, on the contrary, that for some F_1, F_2 ; G_1, G_2 the efficacy of a test is better then that of the best test in the K-class. Then for the case of F_1, F_2 ; G^*, G^* ; (where G^* is easily seen to be a *bona fide* distribution function), their efficacies remain unchanged, i.e., the K-class is still inferior, because now the combination

$$\frac{(1 - G^*)(1 - G^*)}{(1 - \lambda)(1 - G^*) + \lambda(1 - G^*)} = 1 - G^*$$

is the same as before.

But in this case, we have equal censoring in the two samples and it is well known (Gill, 1980, Section 5.3) that the best test in the K-class is most powerful. The contradiction proves our lemma. ■

Using this lemma we have the following

Theorem 8.1 The mean test as proposed in Chapter seven always has Pitman efficacy less than or equal to the optimal test chosen from the K-class.

Proof: Observe that the variance of the difference of the means test (7.1) is σ_A^2 as defined in Theorem 7.1. We easily recognize that its null variance depends on censoring only through the same combination as (8.2). So the efficacy of the difference of means test only depends on G_1, G_2 through G^*. Applying Lemma 1, we see that the theorem is true. ■

The difference of means test, although less powerful then the optimal one in the K-class, has the feature of "not needing to choose a weight function". However, using the K-class of tests, we always face the problem of choosing the weight function. For example, the Mantel-Haenszel test, which is the optimal test in the K-class assuming constant hazard ratio alternative, behaves very badly in some cross hazard alternative situations as we see in Table 2.

Theorem 8.2 If the weight function J of the rank statistics (3.3) is strictly monotone, then the test is consistent for stochastically ordered alternatives, provided $\int_{T_{12}}^{\infty} J(H)dF_1 = o_P(\frac{1}{\sqrt{N}})$.

Proof: Notice that

$$\int_0^{\infty} J(\lambda F_1 + (1-\lambda)F_2\,)dF_1 - \int_0^1 J(s)ds \neq 0 \quad ,$$

because J is strictly monotone and because of the stochastic order of F_1 and F_2 .

Therefore, we have

$$\sqrt{N}\,(\int_0^{T_{12}} J(\lambda\hat{F}_1 + (1-\lambda)\hat{F}_2\,)d\hat{F}_1 \; - \; \int_0^1 J(s)ds\,)$$

$$= \sqrt{N}\,(\,\int_0^{T_{12}} J(\lambda\hat{F}_1 + (1-\lambda)\hat{F}_2\,)d\hat{F}_1 \; - \; \int_0^{\infty} J(\lambda F_1 + (1-\lambda)F_2)dF_1\,)$$

$$+ \; \sqrt{N}(\,\int_0^{\infty} J(\lambda F_1 + (1-\lambda)F_2)dF_1 \; - \; \int_0^1 J(s)ds\,) \quad .$$

The first term is asymptotically normally distributed according to Theorem 3.1 of Chapter three, and the second term clearly goes to ± ∞. This means that the test is consistent. ■

Theorem 8.3 The rank statistic as introduced in Chapter three is asymptotically most powerful in the equal censoring situations if we chose the right weight function J . In the case of unequal censoring, this rank statistic (with the right choice of J) has the same efficacy as the best test in the K-class under the same situation, provided $\lim\limits_{u \to 1} \dfrac{D'(u)(1-u)+D(u)}{g(u)} \dfrac{D(u)}{1-u} = 0$.

Proof: We first establish some relationships between

$$\frac{dF^{\mu}(t)}{d\mu}\Big|_{\mu=\mu_0} = D(F_0(t)) \quad \text{and} \quad \lim \sqrt{N}\Big(\frac{\Lambda_{\frac{1}{\sqrt{N}}}(t)}{\Lambda_0(t)} - 1\Big) = \gamma(t) \quad .$$

Without loss of generality, we can assume $\mu_0 = 0$, so that $F^0 = F_0$. Notice that $F = 1 - e^{-\Lambda}$, we have

$$\begin{aligned} F^{\mu}(t) - F^{0}(t) &= 1 - e^{-\Lambda_{\mu}(t)} - [1 - e^{-\Lambda_0(t)}] \\ &= e^{-\Lambda_0(t)} - e^{-\Lambda_{\mu}(t)} = e^{-\Lambda_{\xi}(t)}[\Lambda_{\mu}(t) - \Lambda_0(t)] \\ &= e^{-\Lambda_{\xi}(t)}\int_0^t d\Lambda_{\mu}(s) - d\Lambda_0(s) = e^{-\Lambda_{\xi}(t)}\int_0^t \Big(\frac{d\Lambda_{\mu}(s)}{d\Lambda_0(s)} - 1\Big) d\Lambda_0(s) \quad , \end{aligned}$$

where Λ_{ξ} is between Λ_0 and Λ_{μ} . By taking the limits we see that

$$\frac{dF^{\mu}(t)}{d\mu} = D(F_0(t)) = [1 - F_0(t)]\int_0^t \gamma \, d\Lambda_0(s) \quad ,$$

$$d\frac{D(F_0)}{1 - F_0} = \gamma \, d\Lambda_0 \quad \text{and} \tag{8.3}$$

$$\gamma(t) = [1 - F_0(t)]\Big(\frac{D(u)}{1-u}\Big)'\Big|_{u = F_0(t)} \quad .$$

Under the sequence of alternatives $F_{A_N} = F^{\frac{1}{\sqrt{N}}}$, we have that the K-class of tests W as defined by Gill (1980 p. 46)

$$\sqrt{N}\, W \overset{D}{\rightarrow} N\Big(-\int_0^{\infty} K\gamma \, d\Lambda_0 \, , \; \int_0^{\infty} K^2 \frac{1}{\lambda_1\lambda_2} \frac{\lambda_1 H_1 + \lambda_2 H_2}{H_1 H_2} d\Lambda_0\Big) \quad ,$$

where $H_i = (1 - G_i)(1 - F_0)$.

So the efficacy is

$$\frac{\left[\int_0^{\infty} K\gamma \, d\Lambda_0\right]^2}{\int_0^{\infty} K^2 \frac{\lambda_1 H_1 + \lambda_2 H_2}{\lambda_1 \lambda_2 H_1 H_2} d\Lambda_0} \quad .$$

Gill (1980) has shown that in this case the best choice of the weight function K is

$$K_{opt} = \frac{H_1H_2}{\lambda_1H_1 + \lambda_2H_2}\gamma \quad ,$$

and so the best efficacy in this case is

$$\lambda_1\lambda_2\int_0^\infty \gamma^2\frac{H_1H_2}{\lambda_1H_1 + \lambda_2H_2}d\Lambda_0(t)$$

$$= \lambda_1\lambda_2\int_0^\infty \gamma^2\frac{(1 - G_1)(1 - G_2)}{\lambda_1(1 - G_1) + \lambda_2(1 - G_2)}dF_0(t)$$

$$= \int_0^\infty \frac{\gamma^2(t)}{g(F_0(t))}dF_0(t) \quad ,$$

where the equality follows by the fact that in our setting of the random censorship model $H_i = (1 - G_i)(1 - F_0)$ and from the definition of $g(F_0)$.

Now, by the relation (8.3) we see the above efficacy can be written as

$$\int_0^1\left[\frac{D'(u)(1 - u) + D(u)}{1 - u}\right]^2\frac{du}{g(u)} \quad . \tag{8.4}$$

On the other hand, the best rank statistic of Chapter three has efficacy (substitute the optimal J' (4.4) and (4.5) in its efficacy expression (4.3))

$$\frac{\left[\int_0^1\frac{D(u)}{1 - u}d[\frac{D'(u)(1 - u) + D(u)}{g(u)}]\right]^2}{\int_0^1(\frac{D'(u)(1 - u) + D(u)}{g(u)})^2g(u)\frac{du}{(1 - u)^2}} \quad . \tag{8.5}$$

Now integration by parts in the numerator of (8.5) and noting that $D(0) \equiv 0$ and the assumption we have made, we have

$$\frac{D'(u)(1 - u) + D(u)}{g(u)}\frac{D(u)}{1 - u}\Big|_0^1 = 0 \quad .$$

We see that (8.5) becomes

$$\int_0^1\left[\frac{D'(u)(1 - u) + D(u)}{1 - u}\right]^2\frac{du}{g(u)} \quad , \tag{8.6}$$

which is the same as (8.4).

Because we already know the optimal K-class test is most powerful in the equal censoring case, we see that the theorem has been proved. ■

CHAPTER 9

A SUMMARY

In the two sample testing problem with censored data, there are not many test procedures available other then the K-class of tests. Although the Mantel-Haenszel test (a commonly used K-class test) behaves well in many situations, there is sometimes a need for new test procedures. The median test as studied by Brookmeyer and Crowley (1982) is certainly a possible choice. We hope that the rank test and difference of means test proposed in this thesis will serve as possible candidates outside the K-class of tests. In fact, the simulation results as shown in Chapter 8 favor the difference of means test. Of course, further simulation and perhaps theoretical work is needed before we know when these tests are better than the K-class of tests.

Both the rank test and the difference of means test as proposed in the previous chapters could be generalized in a number of ways. The difference of means test can be generalized to the multi-sample case and, more important, to the linear regression case. In fact, the two and multi-sample problems are special cases of linear regression. Thus techniques similar to those used in this thesis can yield some results for the 'synthetic' data regression problem for censored data. (see e.g., Leurgans 1984). We plan to investigate this problem in future work. On the other hand, it is interesting to generalize the difference of means test to sequential analysis and compare it with the Mantel-Haenszel test in both proportional and nonproportional alternatives.

For the rank test, the proof of Theorem 3.1 with a little extra care could allow the statistics to 'stop' at some early point (so called administrative censoring) as mentioned in the Remark 3.1. Hence, one can investigate this rank test if multiple looks at the data are planned, i.e., one could study the performance of the test statistics for sequential or group sequential analysis of the data. The weight function J in the (3.3) could be taken as J_N instead of a single one, i.e., choose the weight function adaptively. Generalizations of the rank test to the multi-sample case are also possible.

More simulation study is needed in a variety of possible settings, but that was not the main focus of this thesis due to the limit of time. For instance, it would be interesting to see how the difference of means test would compare to the Mantel-Haenszel test in other situations (heavy and medium tail cases), and the early stopping of the rank test in the nonproportional hazards settings.

REFERENCES

AALEN, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6,** 701-726.

ANDERSEN, P. K. and BORGAN, O. (1985). Counting process models for life history data: A review (with discussion). *Scand. J. Statist.* **12,** 97-158.

ANDERSEN, P. K., BORGAN, O., GILL, R. and KEIDING, N. (1982). Linear nonparametric tests for comparison of counting processes, with applications to censored survival data. *Int. Stat. Rev.* **50,** 219-258.

BILLINGSLEY, P. (1968). *Weak Convergence of Probability Measures.* Wiley, New York.

BRESLOW, N. and CROWLEY, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.* **2,** 437-453.

BROOKMEYER, R. and CROWLY, J. (1982). A k-sample median test for censored data. *J. Amer. Statist. Assoc.* **77,** 433-440.

CHERNOFF, H. and SAVAGE, R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Ann. Math. Statist.* **29,** 972-994.

CHOW, Y. S. and TEICHER, H. (1978). *Probability Theory: independence, interchangeability, martingales.* Springer, New York.

COX, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc. B* **34,** 187-220.

EFRON, B. (1967). The two sample problem with censored data. *Proc. Fifth Berkeley Symp. Math. Statist. Probability* **IV,** 831-853. Univ. California Press.

FOLDES, A. and REJTO, L. (1980). A LIL type result for the product limit estimator. *Z. Wahrsch. verw. Gebiete* **56,** 75-86.

GEHAN, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52,** 203-223.

GELFAND, I. M. and FOMIN, S. V. (1963). *Calculus of Variations.* Prentice-hall, Englewood Cliffs, New Jersey.

GILL, R. (1980). *Censoring and Stochastic Integrals.* Mathematical Centre Tracts 124. Mathematisch Centrum, Amsterdam.

GILL, R. (1983). Large sample behavior of the product-limit estimator on the whole line. *Ann. Statist.* **11,** 49-58.

HARRINGTON, D. P. and FLEMING, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69,** 553-566.

JACOBSON, M. (1982). *Statistical Analysis of Counting Processes.* Lecture Notes in Statistics. Springer, New York.

KALBFLEISCH, J. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data.* Wiley, New York.

KAPLAN, E. and MEIER, P. (1958). Non-parametric estimator from incomplete observations. *J. Amer. Statist. Assoc.* **53,** 457-481.

KOUL, H., SUSARLA, V. and VAN RYZIN, J. (1981). Regression analysis with randomly right-censored data. *Ann. Statist.* **9,** 1276-1288.

KOZIOL, J. A. and GREEN, S. B. (1976). A Cramer-Von Mises statistic for randomly censored data. *Biometrika* **63,** 465-474.

LAI, T. L. (1975). On Chernoff-Savage statistics and sequential rank tests. *Ann. Statist.* **3,** 825-845.

LEURGANS, S. (1984). Linear models, random censoring and synthetic data. *Unpublished manuscript.*

MANTEL, N. (1966). Evaluating of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* **50,** 163-170.

MANTEL, N. and HAENSZEL, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22,** 719-748.

MAURO, D. (1985). A combinatoric approach to the Kaplan-Meier estimator. *Ann. Statist.* **13,** 142-149.

MILLER, R. G. (1981). *Survival Analysis.* Wiley, New York.

PETERSON, A. V. (1977). Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *J. Amer. Statist. Assoc.* **72,** 845-858.

PETO, R. and PETO, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *J. R. Statist. Soc. Ser. A* **135,** 185-206.

PHADIA, E. G. and VAN RYZIN, J. (1980). A note on convergence rates for the product limit estimator. *Ann. Statist.* **8**, 673-678.

PRENTICE, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65**, 167-179.

SUSARLA, V., TSAI W. Y. and VAN RYZIN, J. (1984). A Buckley-James-type estimator for the mean with censored data. *Biometrika* **71**, 624-625.

SUSARLA, V. and VAN RYZIN, J. (1980). Large sample theory for an estimator of the mean survival time from censored samples. *Ann. Statist.* **8**, 1002-1016.

TARONE, R. E. and WARE, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64**, 156-160.

ZHENG, Z. (1984). Regression with randomly censored data. *Ph. D. thesis.* Columbia University.