AN INVESTIGATION INTO THE POWER OF THREE TESTS USED TO COMPARE SURVIVAL DISTRIBUTIONS

by

JOHN STEVEN GATSCHET

B.A., Saint Louis University, 1985

A MASTER'S REPORT

submitted in partial fullfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1987

Approved by:

James J. Higgins

Major Professor

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

ACKNOWLEDGEMENTS

I would like to thank Dr. James Higgins for the guidance and invaluable insights he gave me throughout the formation of this report. In addition, I thank Dr. Paul Nelson and Dr. James Neill for their critical review of this report and for agreeing to sit on my committee. My heartfelt gratitude goes to Retha Parker for transforming the scribbling of a weary hand into the printed word.

I extend a special thanks to the people in the Department of Statistics for two most memorable years. I thank my brothers and sisters, Mark, Denise, Ramona, and Jim, for all of their encouragement. Finally, I thank my parents, Paul and Carolyn. This report is yet another testimony to their constant love and support.

I. INTRODUCTION

A variety of tests exists to compare two distributions. Of interest among these tests are three in particular designed to compare survival distributions. These tests are Gehan's generalized Wilcoxon test, the logrank test, and the likelihood ratio test.

The power of these tests to distinguish between two survival distributions will be investigated in the presence of various factors. These factors include sample size, the level of censoring, and the means of the two distributions. The type of censoring considered in this study is random censoring. Exponential distributions will be used to generate the samples.

In addition to comparing the three tests, a model will be constructed for each respective test to provide a method of predicting the power of the test given a specific combination of the above factors. Of interest here is not only the models themselves, but also the techniques employed to develop them.

Overriding the objectives just delineated is the objective to obtain these results in a manner that is both efficient and accurate while also being applicable to other problems involving powers of tests.

II. DESCRIPTION OF TESTS

A. Gehan's Generalized Wilcoxon Test

The test statistic used here is computed by a procedure developed by Mantel. Suppose we observe N failure times from two survival distributions. Of these N failure times, r are uncensored. Mantel's procedure pools the two samples and ranks this pooled sample in ascending order while ignoring censoring.

For each observation t_i compute

V_i = the number of observations definitely less than t_i

- the number of observations definitely greater than t_i.

Add the V_i corresponding to the first sample (or second sample) to obtain a sum statistic T. T is then used to compute a test statistic

$$W = \frac{(T - E(T))^2}{Var(T)}$$

As discussed in Lee, if the distributions are identical,

$$E(T) = 0$$

and

$$Var\ (T) = \frac{(N - n)n}{N-1} \quad \frac{\Sigma V_i^2}{N}$$

where n denotes the size of sample one and N is the total sample size. The test statistic simplifies to

$$W = \frac{T^2}{Var(T)}$$

and follows a chi-square distribution. If W exceeds the 100α percentage point of a chi-square distribution with one degree freedom, then the hypothesis of no difference in the survival distributions is rejected.

B. Logrank Test

To compute the logrank test statistic the following quantities are needed for all uncensored observations, t_i:

R_i = number of observations surviving and uncensored just before t_i

E_{ik} = the proportion of these R_i in the k^{th} group (k = 1 or 2) .

Then calculate $E_k = \sum_i E_{ik}$

O_k = observed number of failures in the k^{th} group.

The test statistic L is computed as

$$L = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} .$$

This statistic follows a chi-square distribution with one degree freedom. As with Gehan's test, reject the hypothesis of no difference between the two distributions if L exceeds the 100α percentage point of a chi-square distribution.

C. Likelihood Ratio Test

Suppose there are n_1 and n_2 survival times in groups one and two, respectively. In group one r_1 values are uncensored, and $n_1 - r_1$ values are censored. Likewise, in group two r_2 values are uncensored and $n_2 - r_2$ values are censored. Let T_1 be the sum of survival times (censored and uncensored) from the first group, and let T_2 be the sum of survival times from the second group.

Since the distributions involved in this study will be exponential distributions with density $f(t) = \lambda e^{-\lambda t}$, $t \geq 0$, testing the equality of distributions is equivalent to testing

$$H_0: \lambda_1 = \lambda_2 = \lambda$$

$$H_a: \lambda_1 \neq \lambda_2$$

where λ_1 and λ_2 are 1/mean for the two respective distributions.

The test statistic R is a ratio of two likelihood functions

$$R = \frac{L(\hat{\lambda}, \hat{\lambda})}{L(\hat{\lambda}_1, \hat{\lambda}_2)} .$$

The numerator is the maximized logarithm of the likelihood function for the combined sample under the null hypothesis. The denominator is the maximized logarithm of the likelihood function for the two groups combined. The quantity $\hat{\lambda}$ is the maximum likelihood estimate for λ namely

$$\hat{\lambda} = \frac{r_1 + r_2}{T_1 + T_2} .$$

The quantities $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are maximum likelihood estimates of λ_1 and λ_2,

$$\hat{\lambda}_1 = \frac{r_1}{T_1}$$

and

$$\hat{\lambda}_2 = \frac{r_2}{T_2} .$$

The values of $L(\hat{\lambda},\hat{\lambda})$ and $L(\hat{\lambda}_1, \hat{\lambda}_2)$ are then computed as

$$L(\hat{\lambda},\hat{\lambda}) = \hat{\lambda}^{r_1+r_2} \exp(-r_1-r_2)$$

and

$$L(\hat{\lambda}_1,\hat{\lambda}_2) = \hat{\lambda}_1^{r_1}\hat{\lambda}_2^{r_2} \exp(-r_1 - r_2).$$

R is then calculated as a ratio of these two quantities. The quantity-2logR has an approximate chi-square distribution with one degree of freedom under the null hypothesis. Thus the hypothesis of no differences between the two survival distributions is rejected at level α if -2logR exceeds the 100α percentage point for the chi-square distribution with one degree of freedom.

D. Numerical Example

Suppose the following survival times were observed.

Table 1. Sample survival times.

Group	Time	Group	Time
1	0.34	2	0.38
1	0.46+	2	0.77+
1	0.50	2	0.85
1	1.67+	2	0.96
1	1.79	2	1.23+
1	2.04	2	3.45+

+ = censored observation

Gehan's generalized Wilcoxon test would be performed as follows:

Group	Time	V_i
1	.34	0 - 11 = - 11
2	.38	1 - 10 = - 9
1	.46+	2 - 0 = 2
1	.50	2 - 8 = - 6
2	.77+	3 - 0 = 3
2	.85	3 - 6 = - 3
2	.96	4 - 5 = - 1
2	1.23+	5 - 0 = 5
1	1.67+	5 - 0 = 5
1	1.79	5 - 2 = 3
1	2.04	6 - 1 = 5
2	3.45+	7 - 0 = 7

$$T = -11 + 2 - 6 + 5 + 3 + 5 = -2$$

$$\mathrm{Var}(T) = \frac{6(6)}{11} \cdot \frac{394}{12} = 107.45$$

$$W = \frac{T^2}{\mathrm{Var}(T)} = \frac{4}{107.45} = .04$$

Compare this to a chi-square critical value of 3.84 and conclude there is no significant difference between the two survival distributions.

The logrank test yields the following results:

Time	R_i	E_{i1}	E_{i2}
0.34	12	6/12	6/12
0.38	11	5/11	6/11
0.46+	--	--	--
0.50	9	4/9	5/9
0.77+	--	--	--
0.85	7	3/7	4/7
0.96	6	3/6	3/6
1.23+	--	--	--
1.67+	--	--	--
1.79	3	2/3	1/3
2.04	2	1/2	1/2
3.45+	--	--	--

$$E_1 = \Sigma E_{i1} = 3.49$$

$$E_2 = \Sigma E_{i2} = 3.51$$

$$O_1 = 4 \text{ and } O_2 = 3 \ .$$

Hence,

$$L = \frac{(4 - 3.49)^2}{3.49} + \frac{(3 - 3.51)^2}{3.51} = .149$$

Compare this to a chi-square critical value of 3.84, and conclude there is no significant difference between the two survival distributions.

The likelihood ratio test yields the following results:

$$T_1 = 0.34 + 0.46 + 0.50 + 1.67 + 1.79 + 2.04 = 6.80$$

$$T_2 = 0.38 + 0.77 + 0.85 + 0.96 + 1.23 + 3.45 = 7.64$$

$$r_1 = 4 \text{ and } r_2 = 3.$$

Thus

$$\hat{\lambda}_1 = 4/6.80$$

$$= .5882$$

$$\hat{\lambda}_2 = 3/7.64$$

$$= .3927$$

and

$$\hat{\lambda} = \frac{4 + 3}{6.80 + 7.64} = .4848 \quad .$$

Hence,

$$L(\hat{\lambda},\hat{\lambda}) = (.4848)^7 \exp(-7) = 5.736 \times 10^{-6}$$

$$L(\hat{\lambda}_1, \hat{\lambda}_2) = (.5882)^4(.3927)^3 \exp(-7) = 6.610 \times 10^{-6}.$$

Therefore

$$R = \frac{5.736 \times 10^{-6}}{6.610 \times 10^{-6}} = .8678$$

and

$$-2\log R = 0.284.$$

Compare this quantity with a chi-square critical value of 3.84 and conclude that there is no significant difference between the two survival distributions.

III. SIMULATION

A. Factors

The data for this study were obtained through a computer simulation. Many factors were involved in the generation of the data. As mentioned previously, all samples were derived from exponential distributions. The exponential distribution is a reasonable choice for generating survival data since many survival data exhibit exponential behavior.

To investigate the power of the tests in question various combinations of three factors (or treatments) were considered. These factors were sample size, the level of censoring, and the relative means of the two samples.

The two sample sizes considered in this study were ten and twenty. Samples of size ten were generated from one exponential distribution and then compared to a sample size ten generated from a second distribution. Each test was then applied to these samples to determine if a significant difference between the two underlying distributions could be detected. This was then repeated for samples of size twenty.

A second factor which determined the makeup of the samples to which the tests were applied was the level of censoring. The three levels of censoring considered in this study were 0%, 20%, and 40% censoring. What is meant by "level of censoring" is the probability a given observation is censored, namely, 0, .2, and .4, respectively. As mentioned earlier, the type of censoring investigated was random censoring. A censoring distribution was involved in the data generation in addition to the distribution creating the uncensored observations. This censoring distribution was also an exponential distribution.

The level of censoring determined the mean of the censoring distribution. Appendix A shows the derivation for the mean value of the censoring distribution corresponding to the particular level of censoring desired. Thus for a censoring level of 0% no censoring distribution was used. For a 20% level of censoring the mean of the censoring distribution was set to be 4 times the mean of the distribution generating uncensored times. At a level of 40% censoring the mean of the censoring distribution was set to be 3/2 that of the distributional mean associated with the uncensored times.

A given sample was constructed in the following manner. One observation was generated from the censoring distribution, and one observation was generated from another distribution with a fixed mean. These two observations were compared, and the smaller of the two was recorded as the observed survival time. If the censoring distribution generated the smaller observation, the recorded survival time was considered a censored observation. Otherwise, the survival time was considered uncensored.

The final factor affecting the makeup of any two compared samples was their distributional means. One expects power, which is the probability of determining a difference between the two survival distributions, to be equal to the level of significance when the two distributions have identical means and to increase (eventually to one) as the ratio of the second distribution mean to the first increases.

Since one of the objectives in this study is to develop a model for the power function of each test, it was necessary to use means which would yield values of the power function ranging from zero to one. With this in mind, a pilot study (Appendix B) was performed to determine what values for the means should be used with various sample sizes and levels of censoring which would result in the entire continum of the power function. In this pilot study (as in the main simulation) the mean for the first survival distribution was fixed at the value one, and only the second survival distribution's mean was varied. The results of the pilot study led to the choice of 1, 1.5, 2, 3, 4, and 5 as the mean values for the second distribution to be used in the main simulation.

Let us now summarize the simulation to this point. All distributions involved are exponential. The three factors (and corresponding levels) under study are sample size (2), the level of censoring (3), and the ratio of the survival distribution mean (6). Thus there are 2 x 3 x 6 = 36 different factor combinations for which samples are needed.

Since this is an investigation into the power of certain tests, repeated applications of the tests are required for each factor combination. These iterations are needed to measure how often a given test will distinguish differences between two survival distributions given a specific combination of factors. The number of iterations for each test at every factor combination was chosen to be thirty.

The decision to use thirty iterations was made primarily in view of cost. Typically, in studies involving the power of a test, 400 to 1000 applications of the test are made for every treatment combination. The cost to use this number of iterations for this study would be prohibitive. The computer cost for this study using 30 iterations of each test for each treatment combination was roughly 150 dollars. This figure includes generation of the treatment combinations and samples, application of the tests, analysis of variance procedures, calculation of means, and additional programs needed to setup and check the final programs. Left out of this figure of 150 dollars is the cost of model building which should be impervious to the number of test iterations since it deals with power values. If this study were to use 400 iterations of each test for every treatment combination, the cost to perform the identical procedures would be approximately 2,000 dollars. If 1000 iterations of each test were used, the cost would be approximately 5,000 dollars.

To avoid such costs and yet retain an acceptable degree of accuracy in the findings of the study, attention was paid to the manner in which samples were generated. Every sample was generated independently for every factor combination and every iteration within that combination. This was done in hopes of making results additive over any given factor. Thus, for example, consider the factor of sample size set at the level 10. If the number of iterations were added over the factors of censoring and mean for the second distribution, 540 iterations (3 x 6 x 30) of each test were performed for the case in which the sample size was ten.

B. Results of Simulation

SAS was used to perform the simulation. The DATA statement was employed to generate the treatment combinations and their associated samples. The procedure SURVTEST was invoked to perform Gehan's generalized Wilcoxon test, the logrank test, and the likelihood ratio test on the samples.

The result for each test yielded by the SURVTEST procedure was a p-value based on a chi-square distribution with one degree freedom. This p-value is the observed level of significance for a test of the null hypothesis that no differences exist between the two survival distributions. The SAS program implemented to generate the p-values and their associated treatment combinations is given in Appendix C.

IV. COMPARISON OF TESTS

A. Analysis of Power

The first attempt to determine which test was most powerful and in which situations involved an examination of power values. The power for every test was determined in each of the 36 possible treatment combinations. For each treatment combination the power of a test was found by counting the number of p-values among the thirty iterations which were less than or equal to .05. The results of this count are given in Appendix D.

In Appendix D it is seen that the likelihood ratio test detected the most differences in every situation in which the means for the distributions which generated the samples were different, i.e., when the mean for the secsond survival distribution was 1.5, 2, 3, 4, or 5.

This result was true regardless of the sample size or the level of censoring involved. From this analysis we would conclude that the likelihood ratio test is the most powerful of the three tests, at least when dealing with survival data from exponential distributions. Further analysis is needed to determine the more powerful of Gehan's test and the logrank test since the power values presented in this form indicate no clear winner.

One approach considered to determine the more powerful test between Gehan's test and the logrank test involved an investigation of the mean power values. This investigation is meaningful due to the independent manner in which the data were generated. In the figure below are given the mean power values for Gehan's test and the logrank test for each combination of sample size and censoring. These quantities were obtained by computing the average of the six power values (one for every mean of the second distribution) associated with each combination of sample size and censoring.

Table 2. Mean Power Values for Gehan's Test and the Logrank Test

S	C	Gehan's	Logrank
10	0	.4833	.4722
	20	.3500	.3611
	40	.2500	.2611
20	0	.5500	.5833
	20	.4944	.5444
	40	.4555	.4833

S = sample size
C = level of censoring(%)

This table seems to indicate that the logrank test is the more powerful especially for larger sample sizes or when dealing with censored observations.

Another result of this section in need of mention is the probability of Type I error associated with each test. A Type I error is a conclusion stating the two survival distributions are different when in fact they are the same. The observed probability of Type I error is given in Appendix D whenever the mean for the second survival distribution is one since the mean for the first survival distribution is fixed at one. These values are repeated in the table below for the various combinations of sample size and censoring.

Table 3. Observed Probabilities of Type I Error

S	C	Gehan's	Logrank	Likelihood Ratio
10	0	.0667	.0333	.0667
	20	.0000	.0000	.0333
	40	.0000	.0000	.0333
20	0	.0667	.0333	.0667
	20	.0000	.0000	.0333
	40	.0333	.0000	.0000

S = sample size
C = level of censoring (%)

We want these probabilities to be near .05. Since the number of iterations is only thirty for each factor combination, we are unable to tell if we are testing at the .05 significance level. Furthermore, it may be unfair to compare the tests on the basis of power since the significance level may not be the same for each. The way to answer these concerns is to perform a large number of iterations, e.g. 1000, and determine what the actual probabilities of Type I error are and then make some adjustment for differences. The cost to do this using SAS would be prohibitive. However, the models for the power functions developed in Section V of this report show the predicted significance level of each test to be much closer to one another than the observed values given above. Thus, the concerns listed above may not be of too grave a nature.

B. Splitplot Analysis

Another analysis which was done to validate the previous findings was analysis of variance for a splitplot design. Before discussing the results of this splitplot analysis some questions regarding the assumptions involved need to be addressed.

The assumptions needed to perform the splitplot analysis are independence between observations, normally distributed responses, and equal variance among the treatment combinations. Independence is assured by the manner in which the data were generated. The p-values were obtained from tests applied to samples generated independently of one another, so independence is a valid assumption. However, the assumptions of normality and equal variances are not so easily dismissed.

To answer the question of normality a variety of responses were examined. One response investigated was the p-value considered as an integer, i.e. the four-decimal p-value multiplied by 10,000. Another response was the natural logarithm of this integer p-value. In addition, the logarithm of the logarithm of the integer p-value multiplied by ten was investigated. (This multiplication was needed to avoid taking the logarithm of zero since some integer p-values were one.) Finally, the arcsine of the square root the decimal p-value was examined. Another possible response to investigate is the logarithm of p over 1 - p. This transformation was not suited for this study since p-values of size one were observed.

The SAS procedure UNIVARIATE was applied to each of these responses within each of the 36 combinations of sample size, censoring, and distributional mean. The results indicated that out of the 36 possible cells 18 exhibited features of a random sample taken from a normal distribution for the response obtained by taking the natural logarithm of the integer p-value. This was the highest number of such cells observed for any of the responses. (See Appendix E.) Hence, this transformation made by taking the logarithm of the integer p-value was chosen as the response to be used in the splitplot analysis.

In addition, the variance of this response was found to be of similar magnitude for each combination of sample size, censoring, and distributional mean. (See Appendix E.) Hence, the assumptions of normality and equal variance although not strictly satisfied at least were not overtly violated by using the logarithm of the integer p-value as the response variable.

After addressing the questions surrounding the validity of the assumptions, the splitplot analysis was performed. Sample size, the level of censoring, and the mean for the second distribution were designated as the wholeplot treatments. The wholeplot design was completely randomized. The three tests under investigation were the three levels of the subplot treatment. The table below gives the results of the analysis.

Table 4. Splitplot Analysis

Source of Variablilty	df	F	prob > F
S	1	168.3	0.0001
C	2	35.8	0.0001
M	5	298.3	0.0001
S*C	2	0.4	0.6420
S*M	5	23.8	0.0001
C*M	10	2.6	0.0041
S*C*M	10	0.2	0.9978
Wholeplot Error	1044	---	----
T	2	426.5	0.0001
S*T	2	5.9	0.0028
C*T	4	0.6	0.6899
M*T	10	40.5	0.0001
S*C*T	4	2.0	0.0941
S*M*T	10	1.5	0.1455
C*M*T	20	0.9	0.6159
S*C*M*T	20	2.0	0.0056
Subplot Error	2088	---	----

S = sample size

C = level of censoring(%)

M = second distributional mean

T = test

In the subplot portion of the table a significant difference between the mean responses for the three tests was detected. This result may, however, be misleading since there were interaction terms (S*T, M*T, and S*C*M*T) which were also significant. Since there may be a confounding of effects, it is best to refrain from claiming a significant difference exists between the mean responses for each test. Had these interactions not been significant, this analysis would have been quite useful in comparing the three tests.

In the wholeplot portion of the table the factors of sample size, censoring, and the mean for the second distribution are all significant. This indicates that at least two mean responses for each of these factors are significantly different. It is not necessarily inappropriate to make this claim even though some wholeplot interaction terms are significant because we would expect the factors themselves to yield significantly different responses for different levels. These findings suggest that certain trends may exist for the factors among the responses obtained for each test. In the next section an attempt to discover these trends will be made.

C. Trends Among the P-values

As reported in the previous section significant trends exist among the mean responses for various combinations of the factors. These responses involved taking the logarithm of the p-value when considered as an integer. The purpose of this section is not to establish the significance of these trends for the p-values themselves but to give the reader a clearer picture of how the p-values associated with the various tests are affected by sample size, the level of censoring, and the ratio of the two distributional means.

The following table lists the mean p-value of each test for each of the factors mentioned above. (The mean p-value for each test involving the combination of all these factors is given in Appendix F.)

Table 5. Mean P-value of Each Test for Various Factors

Factor	Level	Gehan's	Logrank	Likelihood Ratio
S	10	.2518	.2399	.2191
	20	.2073	.1970	.1845
C	0	.1895	.1786	.1636
	20	.2314	.2176	.2090
	40	.2677	.2591	.2326
M	1	.5299	.5483	.5368
	1.5	.3668	.3492	.3328
	2	.2592	.2305	.2054
	3	.1138	.1028	.0848
	4	.0647	.0485	.0277
	5	.0429	.0315	.0230

S = sample size

C = level of censoring (%)

M = second distributional mean

From this table we can make the following statements. For each test procedure the mean p-value decreases as sample size increases. In addition, as the level of censoring increases, the mean p-value for each test also increases. And finally, as the ratio the two distributional means increases (or as the mean for the second survival distribution increases) the mean p-value for each test decreases. These mean p-values suggest that the likelihood ratio test is the more poweful test to the extent that a lower p-value indicates more power.

D. Analysis of Differences

A way of skirting the difficulty faced in interpreting the splitplot analysis is to redefine the response function in such a manner that the subplot treatment structure is eliminated. To achieve this the following differences were defined:

$$\text{Diff1} = \log(p_1) - \log\ (p_2)$$

$$\text{Diff2} = \log(p_1) - \log(p_3)$$

$$\text{Diff3} = \log(p_2) - \log(p_3)$$

where p_1 is the integer p-value given by Gehan's test; p_2 is that given by the logrank test; and p_3 is the integer p-value given by the likelihood ratio test. Since the $\log(p_i)$ was approximately normally distributed, the difference between two such quantities will also be approximately normal.

An analysis of variance using a completely randomized design was performed for each of the differences defined above. Before investigating these results a few comments are in order. The response under investigation is the difference between the logarithms of two integer p-values each associated with a different test. These p-values should be uniformly distributed over the interval 0 to 1 when the means for the two survival distributions are identical i.e. both one. Furthermore, the p-values should be very near zero when the ratio of the second distributional mean to the first is large. For this study the smallest p-value possible is .0001. (This limitation is made by SAS and not by the test procedures themselves.) Thus, since the previous comments are true regardless of the test procedure involved, the difference between the logarithms of two integer p-values associated with two test procedures is not likely to be significant when the second distribution mean is 1 or quite large (perhaps 5). Therefore, the focus of this analysis will be on the differences in the responses from each test that may exist between these two extremes of the second distributional mean and on the factors of sample size and censoring that may be responsible for these differences.

The results of the previous section involving trends among the p-values will be used in the analysis to follow. These results were that for every test the mean p-value decreases as sample size increases; the mean p-value increases as the level of censoring increases; and the mean p-value decreases as the ratio of the two distributional means increases. These same trends will hold true for the logarithm of the integer p-value since the logarithm of a number increases as that number increases.

The analysis of variance table for the response Diffl is given below.

Table 6. Analysis of Variance for the Response Comparing Gehan's Test and the Logrank Test.

Source of Variablilty	df	F	prob > F
S	1	25.09	.0001
C	2	3.26	.0389
M	5	20.13	.0001
S*C	2	0.18	.8361
S*M	5	1.87	.0953
C*M	10	0.58	.8312
S*C*M	10	1.95	.0360
Error	1044	---	---

S = sample size

C = level of censoring (%)

M = second distributional mean

Recall the response Diffl allows for the comparison of Gehan's test and the logrank test. As seen in the table above, the factors of sample size, censoring, and second distributional mean are all significant.

Focusing on the factor of sample size, we see the mean difference between the responses for each test when the sample size is ten is significantly different than the mean difference between the responses for each test when the sample size is twenty. These two mean differences are (from Appendix G) .1475 when sample size is ten and .3418 when sample size is twenty. The mean difference between the responses is the same as the difference between the means of the responses (response being the logarithm of the integer p-value). Therefore, the mean response for the logrank test decreases more in going from a sample size of ten to a sample size of twenty than the mean response for Gehan's test.

For the factor of censoring we discover another pattern. The mean differences here are .3009, .2526, and .1804 for censoring levels of 0%, 20%, and 40%, respectively. Thus as censoring increases the mean response for Gehan's test increases less quickly than the mean response for the logrank test. However, the mean response for Gehan's test is still larger at every level of censoring.

For the values of the second distributional mean 1, 1.5, 2, 3, 4, and 5 the observed mean differences were -.0565, .0905, .1518, .3791, .4681, .4349, respectively. These results suggest that the mean response for the logrank test decreases more quickly than the mean response for Gehan's test as the mean for the second survival distribution increases. Notice that the mean difference decreases in magnitude when the second distributional mean goes from 4 to 5. This decrease is expected and would continue, eventually becoming quite small, since the responses for both tests are based on p-values which eventually assume the value .0001 as the disparity between the two distributional means becomes quite large. (As mentioned previously, .0001 is the smallest p-value SAS reports for any of the three tests under study.)

Another meaningful analysis to compare Gehan's test to the logrank test was done using the SAS procedure GLM and the LSMEANS statement. In addition to the mean differences used above, the LSMEANS statement with the STDERR option gave the significance level of a t-test comparing each mean difference to zero. These results are found in Appendix G. If a mean difference is significantly different from zero and is a positive quantity, then the mean response for Gehan's test is significantly larger than the mean response for the logrank test (in the case of Diff1).

The table below, taken from Appendix G reports the mean difference between the responses for Gehan's test and the logrank test for the factors of sample size, censoring, and the second distributional mean.

Table 7. Means for the Response Diff1 and the Observed Significance Levels of the Test H_0: mean = 0.

Factor	Level	Mean Difference	Prob > T
S	10	.1475	.0001
	20	.3418	.0001
C	0	.3009	.0001
	20	.2526	.0001
	40	.1804	.0001
M	1	-.0565	.2344
	1.5	.0905	.0570
	2	.1518	.0014
	3	.3791	.0001
	4	.4681	.0001
	5	.4349	.0001

S = sample size
C = level of censoring (%)
M = second distributional mean

For the factor sample size it is seen that the mean response for Gehan's test is significantly larger than the mean response for the logrank test for samples of size ten or twenty. (Again, the mean response is the natural logarithm of the integer p-value.) The implication of this result is that the logrank test is the preferred test regardless of the sample size involved.

For the three levels of censoring the mean response for Gehan's test again is significantly larger than the mean response for the logrank test. Therefore, regardless of the level of censoring, the logrank test is preferred over Gehan's test.

Finally, for the various values of the second distributional mean the mean response for the logrank test is significantly less than that of Gehan's test for all values except 1 and 1.5. For the value 1 this is expected as mentioned previously. For the value 1.5 it is not surprising since the ratio of the two distributional means is not quite large enough for a significant difference between the mean responses of the two tests to exist.

The mean differences and the associated significance levels for the combination of all three factors are reported in Appendix G. The pattern is that the logrank test has a significantly lower response than Gehan's test in every situation in which the mean for the second survival distribution is not 1 or 1.5. A lower response implies a lower p-value and perhaps a greater ability to detect differences. To the extent that p-values indicate power, the logrank test is the better test regardless of the sample size or the level of censoring involved.

To compare Gehan's test to the likelihood ratio test, the response Diff2 is used. Recall this is the logarithm of the integer p-value for Gehan's test minus the logarithm of the integer p-value for the likelihood ratio test. The analysis of variance table for this response is given below.

Table 8. Analysis of Variance for the Response Comparing Gehan's Test to the Likelihood Ratio Test.

Source of Variablilty	df	F	prob > F
S	1	5.76	.0165
C	2	0.12	.8883
M	5	48.46	.0001
S*C	2	1.78	.1692
S*M	5	1.45	.2008
C*M	10	0.90	.5295
S*C*M	10	1.69	.0784
Error	1044	---	---

S = sample size

C = level of censoring (%)

M = second distributional mean

At the .05 level the significant factors are sample size and the second distributional mean. The mean differences (from Appendix G) associated with samles of size ten and twenty are .8345 and 1.0345, respectively. Thus as sample size increases the mean response for the likelihood ratio test decreases more quickly than the mean response for Gehan's test.

The mean differences associated with the values of 1, 1.5, 2, 3, 4, and 5 for the second distributional mean are -0.0196, 0.3455, 0.6767, 1.2531, 1.6827, and 1.6739, respectively. Thus, as the second distributional mean increases the mean response of the likelihood ratio test decreases more quickly than the mean response of Gehan's test. Notice the slight trend in the opposite direction when the second distributional mean moves from 4 to 5. This trend is expected and will continue since both tests are yielding p-values very near or equal to .0001 because the disparity between the means of the two survival distributions is becoming quite large.

The analysis of variance table does not indicate a significant difference in the mean differences associated with the three levels of censoring.

The results of the LSMEANS statement with the STDERR option applied to the response Diff2 are shown in the table below.

Table 9. Means for the Response Diff2 and the Observed Significance Levels of the Test H_0: mean = 0.

Factor	Level	Mean Difference	Prob > T
S	10	0.8345	.0001
	20	1.0349	.0001
C	0	0.9579	.0001
	20	0.9377	.0001
	40	0.9084	.0001
M	1	-0.0196	.8477
	1.5	0.3455	.0008
	2	0.6727	.0001
	3	1.2531	.0001
	4	1.6827	.0001
	5	1.6739	.0001

S = sample size
C = level of censoring (%)
M = second distributional mean

From this table it is seen that the mean response for the likelihood ratio test is significantly less than the mean response for Gehan's test for all levels of sample size and censoring. Furthermore, the mean response for the likelihood ratio test is significantly less for every value of the second distributional mean except 1, where a significant difference is not expected.

The pattern for the combination of all these factors (given in Appendix G) is that the mean response for the likelihood ratio test is always significantly less except when the second distributional mean is 1 and occasionally 1.5.

Since the mean response for each test is simply a transformation of the p-value given by that test, we conclude that the observed significance level for the likelihood ratio test is, on average, less than that for Gehan's test, regardless of the levels of sample size or censoring.

The response Diff3 was defined to compare the logrank test to the likelihood ratio test. This response is the logarithm of the integer p-value yielded by the logrank test minus the logarithm of the integer p-value yielded by the likelihood ratio test. The results of an analysis of variance for this response are found in the table below.

Table 10. Analysis of Variance for the Response Comparing the Logrank Test to the Likelihood Ratio Test.

Source of Variabiilty	df	F	prob > F
S	1	0.01	.9291
C	2	0.36	.6981
M	5	35.26	.0001
S*C	2	2.86	.0576
S*M	5	1.36	.2377
C*M	10	0.94	.4987
S*C*M	10	2.44	.0070
Error	1044	---	---

S = sample size
C = level of censoring (%)
M = second distributional mean

From this table it is seen that the mean differences for the various values of th second distributional mean were found to be significantly different. However, the mean differences for the various sample sizes were not found to be significant, nor were the mean differences for the three levels of censoring.

The values for the mean differences and the corresponding significance levels of a t-test that these mean differences are zero are given in the table below.

Table 11. Means for the Response Diff3 and the Observed Significance Levels of the Test H_0: mean = 0.

Factor	Level	Mean Difference	Prob > T
S	10	0.6870	.0001
	20	0.6931	.0001
C	0	0.6570	.0001
	20	0.6851	.0001
	40	0.7280	.0001
M	1	0.0369	.6620
	1.5	0.2550	.0026
	2	0.5209	.0001
	3	0.8740	.0001
	4	1.2146	.0001
	5	1.2389	.0001

S = sample size
C = level of censoring (%)
M = second distributional mean

The table indicates that the likelihood ratio test has a significantly smaller mean response than the logrank test for every level of sample size and censoring. This in turn implies (by arguments developed earlier) that the observed significance level for the likelihood ratio test is, on average, less than that of the logrank test for the various levels of sample size and censoring.

Let us now summarize the findings of this section on the analysis of differences. (Recall the effects of sample size and censoring on the responses for each test were of primary interest.) Increasing sample size decreased the mean response for the likelihood ratio test the quickest, followed by the logrank test, and finally Gehan's test. The mean response of the likelihood ratio test was significantly smaller than that of the other tests for both sample sizes. The mean response for the logrank test was significantly less than that for Gehan's test.

Increasing the level of censoring increased the mean response for the logrank test the quickest, followed by the likelihood ratio test, and finally Gehan's test. However, the mean response for Gehan's test was significantly larger than the mean responses for the other two tests at all levels of censoring. The smallest mean responses at all the levels of censoring came from the likelihood ratio test.

The significance of these findings is determined by the strength of the following statements. If the mean response of a test for a given factor is smaller than the mean response for another test, then the observed significance level is also smaller on average. (This is the case because the response analyzed for each test was a simple transformation of the p-value.) To the extent that significance level measures the ability of a test to distinguish differences between two survival distributions, the likelihood ratio test is best, followed by the logrank test, and finally Gehan's test. The particular sample size or level of censoring involved does not affect this result.

V. MODELS FOR THE POWER FUNCTIONS

A. Model Building

The last portion of this study is concerned with building a power function for each of the three tests under investigation. The power function for each test was constructed seperately; however, the manner in which the power function was developed was identical for each test.

The responses used for model building were the 36 power values associated with the 36 combinations of sample size, censoring, and second distributional mean. (See Appendix D). Logistic regression was used to build each power function. The model used in logistic regression is

$$Y = \log(p/1 - p) = \beta_0 + \beta_1 X_1 + .. + \beta_k X_k$$

where p is the value of the power function for a particular configuration of the predictor variables. Estimates for the beta parameters were calculated by the maximum likelihood method. Once these estimates are obtained, a predicted value of the power function for a specific configuration of the predictor variables can be found by

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + .. + \hat{\beta}_k X_k}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + .. + \hat{\beta}_k X_k}} .$$

In this section we wish to develop models which will predict the power of the test for any level of sample size, censoring, and mean for the second survival distrbution. This means that the variables used to model the power function will be considered as continuous variables and not as categorical variables. Hence, the models developed will be able to predict power for levels of the three factors not necessarily observed in this study. However, as with any regression problem, the model might not be accurate for levels outside the ranges of those observed.

The SAS procedure FUNCAT was used to perform the logistic regression. The statement DIRECT was also used in order that the various factors would be considered as continuous variables. To determine the better model between two competing models the following criterion was used. The model which yielded the smallest mean squared residual (MSR) was considered the better model. This was calculated as

$$\text{MSR} = \frac{\sum_{i=1}^{36} (p_i - \hat{p}_i)^2}{36 - k - 1}$$

where k is the number of predictor variables in the model. This criterion was chosen since it considers both the simplicity (number of predictor variables) and the accuracy (size of residuals) in determining the best model. In addition to printing residuals, the FUNCAT procedure does a chi-squared test of the hypothesis $\beta_i = 0$ for each parameter in the model.

The method of model building used in this section is similar to a backward stepwise regression. The first model to be fitted was the complete fourth-order model. This model is

$$\begin{aligned} Y = {} & \beta_0 + \beta_1 S + \beta_2 C + \beta_3 M + \beta_4 S^2 + \beta_5 C^2 \\ & + \beta_6 M^2 + \beta_7 SC + \beta_8 SM + \beta_9 CM + \beta_{10} S^3 + \beta_{11} C^3 \\ & + \beta_{12} M^3 + \beta_{13} S^2 C + \beta_{14} S^2 M + \beta_{15} SC^2 + \beta_{16} SM^2 \\ & + \beta_{17} SCM + \beta_{18} C^2 M + \beta_{19} CM^2 + \beta_{20} S^4 + \beta_{21} C^4 \\ & + \beta_{22} M^4 + \beta_{23} S^3 C + \beta_{24} S^3 M + \beta_{25} S^2 C^2 + \beta_{26} S^2 M^2 \\ & + \beta_{27} S^2 CM + \beta_{28} SC^3 + \beta_{29} SM^3 + \beta_{30} SC^2 M + \beta_{31} SMC^2 \\ & + \beta_{32} C^3 M + \beta_{33} C^2 M^2 + \beta_{34} CM^3 \end{aligned}$$

where S denotes sample size, C is level of censoring, and M is the mean of the second survival distribution. (In this study the term of

S^2 is dropped since it has no degrees of freedom associated with it.) The MSR was then calculated for this model. Next, the term(s) with the highest p-value for the test of $\beta_i = 0$ were eliminated. This new reduced model was then fit and MSR calculated for it. The MSR for this reduced model was then compared to that for the complete fourth-order model. If less, another term was eliminated and a new reduced model was fit. This process continued until no term could be dropped from the model without increasing MSR.

As mentioned previously, this procedure was performed seperately for Gehan's test, the logrank test, and the likelihood ratio test. The following table reports the estimated power function for each of the tests. (Terms without a parameter estimate were not included in the final model.)

Table 12. Estimated Power Functions.

Term	Gehan's	Logrank	Likelihood Ratio
Intercept	-15.7358	-12.6753	-14.7175
S	0.291561	- 2.33946E-2	3.2633E-2
C	- 0.28273	- 5.77564E-2	- 2.82213E-2
M	17.5645	13.4989	17.7001
C^2	7.22364E-3	---	6.63594E-4
M^2	- 7.22562	- 5.74508	- 8.44528
SC	---	- 2.87945E-3	---
SM	- 0.335331	---	---
CM	0.150668	3.32825E-2	---
S^3	---	---	---
C^3	---	---	---
M^3	1.27569	1.07205	1.83239
S^2C	---	2.05939E-4	---
S^2M	---	3.57452E-2	---
SC^2	---	---	---
SM^2	.129754	---	---
SCM	---	---	---
C^2M	- 4.07982E-3	---	- 1.7206E-4
CM^2	- 2.48676E-2	- 1.52535E-2	---
S^4	---	---	---
C^4	---	---	---
M^4	- 7.93742E-2	- 7.02037E-2	- 0.15019
S^3C	---	---	---
S^3M	---	---	---
S^2C^2	---	---	---
S^2M^2	---	---	---
S^2CM	---	---	---
SC^3	---	---	---
SM^3	- 1.44831E-2	- 6.47913E-3	3.18787E-3
SC^2M	- 1.669E-5	- 1.1521E-4	---
SCM^2	3.4407E-4	9.40611E-4	---
C^3M	---	---	---
C^2M^2	5.7975E-4	1.71869E-4	---
CM^3	---	---	---

S = sample size C = level of censoring (%)
M = second distributional mean

To better illustrate these power functions the following table was constructed. Given in the table are predicted and observed values of the power function for each of the tests at various factor combinations. The observed values are those derived from the simulation and found in Appendix D. The predicted values are obtained by putting the values of sample size, censoring, and second distributional mean into the previous prediction equations then solving for $\hat{p}$, as shown earlier.

Table 13. Predicted and Observed Power Values for Various Factor Combinations.

S	C	M		Gehan's	Logrank	Likelihood Ratio
10	0	1	P	.02973	.02062	.03161
			O	.06667	.03333	.06667
10	0	1.5	P	.19361	.14946	.21226
			O	.13333	.10000	.20000
10	20	2	P	.22176	.22840	.33234
			O	.33333	.30000	.36667
10	20	3	P	.49347	.47053	.59048
			O	.43333	.43333	.46667
10	40	4	P	.38600	.45504	.72728
			O	.43333	.43333	.76667
10	40	5	P	.48863	.55618	.66780
			O	.46667	.56667	.66667
20	0	4	P	.90242	.93795	.98918
			O	.93333	.93333	1.00000
20	0	5	P	.90603	.92613	.99843
			O	.90000	.93333	1.00000
20	20	1	P	.01194	.01018	.03131
			O	.00000	.00000	.03333
20	20	1.5	P	.09185	.10757	.21761
			O	.06667	.13333	.20000
20	40	2	P	.25326	.28608	.42621
			O	.20000	.23333	.40000
20	40	3	P	.55106	.62450	.76232
			O	.56667	.63333	.76667

S = sample size
C = level of censoring(%)
M = second distributional mean

P = predicted value
O = observed value

B. Validation of Models

To validate the models developed in the previous section another simulation was done. To limit cost this second simulation looked at two specific combinations of the factors. Both combinations had the factor sample size set at 15 and the level of censoring at 20%. These values were chosen since they fell in the "middle" of the observed ranges. The two values chosen for the second distributional mean were 2 and 3 since these yielded values of the power function which were approximately .25 and .75, respectively, in the first simulation. The following table reports the predicted power values and the power values obtained from the simulation. The power values obtained from the simulation are based on the number of rejections (at the .05 significance level) out of the 100 iterations performed for each factor combination.

Table 14. Predicted and Observed Power Values when Sample Size is 15, Level of Censoring is 20%, and Second Distributional Mean (M) is 2 and 3.

Test		Predicted	Observed
Gehan's	M = 2	.2565	.3100
	M = 3	.6212	.6100
Logrank	M = 2	.3016	.3600
	M = 3	.6852	.7200
Likelihood	M = 2	.3996	.4300
Ratio	M = 3	.7230	.7900

From this table it is seen that the predicted values are relatively close to the observed values. The models for the predicted values were based on thirty iterations of each factor combination while the observed values were based on 100 iterations. If the observed values can be considered as the actual power values, the apparent conclusion is that massive numbers of iterations are not needed to obtain a respectable estimate of the power function.

VI. CONCLUSIONS

The first issue to be resolved in this study was a comparison of three tests used to detect differences in survival distributions. The factors which affected the makeup of the samples from these distributions were sample size, the level of censoring, and the ratio of the two distributional means. After performing an analysis of the power values, a splitplot analysis and an analysis of the differences between the responses for each test, the likelihood ratio test appeared to be the most powerful, followed by the logrank test, and then Gehan's generalized Wilcoxon test. The factors of sample size and level of censoring do not affect this result.

A second issue to be addressed in this study was building power functions for each test. The functions obtained appear relatively accurate in predicting power values for any level of sample size and censoring, and any ratio of the two distributional means (as long as these are kept within the ranges for the factors used to develop the models).

Finally, the issue of efficiency needs to be discussed. This study was undertaken using a small number of iterations (30) for each factor combination when iterations of size 400 to 1000 are normally used. The primary reason behind choosing such a small number of iterations was cost. Nonetheless, the small number of iterations proved quite adequate in determining which testing procedure was more powerful. Each analysis performed indicated the same results.

With regard to building models for the power functions, 30 iterations was sufficient in getting a relatively close estimate of the actual power value. However, more iterations probably would have improved these estimates. Thus, the overall recommendation is that only a small number of iterations are needed to compare the power of tests and develop models for the power of those tests. However, the accuracy of these power function models has room for improvement.

REFERENCES

Lee, Elisa T. (1980). Statistical Methods for Survival Data Analysis, Belmont, California: Lifetime Learning Publications.

Peto, R. and Pike, M. C. (1973). "Conservatism of the Approximation $\Sigma(0 - E)2/E$ in the Logrank Test for Survival Data or Tumor Incidence Data," Biometrics, 29, 579-584.

SAS Institute Inc. (1982). Sas User's Guide: Basics, Statistics, Supplement, 1982 Edition. Cary, NC: SAS Institute Inc.

APPENDIX

APPENDIX A

Derivation of the mean for the censoring distribution.

Let T and C be two independent random variables distributed exponentially with parameters λ and λ_c, respectively. The joint density function of T and C is:

$$h(C,T) = f_C(C)f_T(T)$$

$$= \lambda_c\lambda\ e^{-\lambda_c C\ -\ \lambda T}$$

Now consider the transformation

$$Y_1 = C$$

$$Y_2 = T - C\ .$$

The Jacobian for this transformation is

$$J = \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} = 1.$$

The joint density of Y_1 and Y_2 is

$$g(Y_1,\ Y_2) = h(C,\ T)\ \cdot\ |J|$$

$$=\ \lambda\lambda_c e^{-Y_1(\lambda\ +\ \lambda_c)\ -\ \lambda Y_2}\ .$$

From this we obtain the marginal density

$$g(Y_2) = \int_{Y_1} g(Y_1,\ Y_2)dY_1$$

$$= \lambda\lambda_c e^{-\lambda Y_2} \int e^{-(\lambda\ +\ \lambda_c)Y_1}dY_1$$

$$= \lambda\lambda_c e^{-\lambda Y_2} \left[-\frac{1}{\lambda + \lambda_c} e^{-(\lambda + \lambda_c)Y_1} \right]_{Y_1}$$

We need only consider this density for $Y_2 > 0$. For this support the density is

$$g(Y_2) = \int_0^\infty g\ (Y_1,\ Y_2)dY_1$$

$$= \frac{\lambda\lambda_c}{\lambda + \lambda_c} e^{-\lambda Y_2} .$$

A observation is censored if the value c from the censoring distribution is less than the value t from the survival time distribution. Hence the probability an observation is censored is

$$P(C < T) = P(0 < T - C)$$

$$= P(0 < Y_2)$$

$$= \int_0^\infty \frac{\lambda\lambda_c}{\lambda + \lambda_c} e^{-\lambda Y_2} dY_2$$

$$= \frac{\lambda_c}{\lambda + \lambda_c} .$$

Let P be the probability an observation is censored. Solving for λ_c in the above equation obtains

$$\lambda_c = \frac{P}{1 - P} \lambda .$$

To put this in terms of the mean of the exponential distributions involved

$$\mu_c = \frac{1 - P}{P} \mu$$

where μ_c is the mean of the censoring distribution and μ is the mean of the distribution of the uncensored survival times.

Thus for the levels of censoring involved in this study, namely, 20% and 40%

$$\mu_c = 4\mu$$

and

$$\mu_c = 3/2\mu,$$

respectively.

APPENDIX B

Pilot study to determine the proper mean values for the second survival distribution.

The values reported are the proportion of iterations out of ten in which a difference was detected between the two distsribution at the .05 significance level. The levels of sample size (S), censoring (C), and the mean of the second survival distribution (M) are also given. The mean for the first survival distribution was fixed at one.

S	C	M	Gehan's	Logrank	Likelihood Ratio
10	0	1.5	0.3	0.2	0.3
		3	0.6	0.9	0.9
		5	1.0	1.0	1.0
	20	1.5	0.1	0.2	0.3
		3	0.6	0.7	0.8
		5	0.7	0.7	0.8
	40	1.5	0.0	0.0	0.1
		3	0.3	0.5	0.6
		5	0.8	0.9	0.9
20	0	1.5	0.3	0.2	0.2
		3	0.9	0.9	0.9
		5	0.9	1.0	1.0
	20	1.5	0.1	0.2	0.2
		3	0.7	0.7	0.9
		5	1.0	1.0	1.0
	40	1.5	0.1	0.2	0.2
		3	0.6	0.7	0.8
		5	0.8	0.9	1.0

APPENDIX C

SAS program used to generate treatment combinations, samples, and p-values.

```
DATA A;
SEED = 876254917;
COUNT = 1;
DO SAMPSIZE = 10,20;
  DO CENLEVL = 0,4,1.5;
    DO MEAN2 = 1,1.5,2,3,4,5;
      DO ITER = 1 TO 30;
        DO SAMPE = 1,2;
          IF SAMPLE = 1 THEN MEAN = 1;
            ELSE MEAN = MEAN2;
          DO I = 1 TO SAMPSIZE;
             T = MEAN*RANEXP(SEED);
             S = CENLEVL*MEAN*RANEXP(SEED)
             IF CENLEVL = 0 THEN TOBS = T;
               ELSE TOBS = MIN(T,S);
             IF TOBS = T THEN CENSOR = 2;
               ELSE CENSOR = 1;
             OUTPUT;
           END;
        END;
        COUNT = COUNT + 1;
      END;
    END;
  END;
END;

PROC SURVTEST;
  BY COUNT;
  CLASS SAMPLE;
  VAR TOBS CENSOR;
```

APPENDIX D

Power values for various treatment combinations.

The values reported are the proportion of iterations out of thirty in which a difference was detected between the two distributions at the .05 significance level. The levels of sample size (S), censoring (C), and the mean of the second survival distribution (M) are also given. The mean for the first survival distribution was fixed at one.

S	C	M	Gehan's	Logrank	Likelihood Ratio
10	0	1	.0667	.0333	.0667
		1.5	.1333	.1000	.2000
		2	.4000	.3667	.4333
		3	.6667	.6667	.7333
		4	.8000	.8000	.8333
		5	.8333	.8667	.9000
	20	1	.0000	.0000	.0333
		1.5	.0333	.1000	.1333
		2	.3333	.3000	.3667
		3	.4333	.4333	.4667
		4	.6000	.6333	.8667
		5	.7000	.7000	.8333
	40	1	.0000	.0000	.0333
		1.5	.1333	.0667	.1333
		2	.2000	.1667	.3667
		3	.2667	.3333	.4333
		4	.4333	.4333	.7667
		5	.4667	.5667	.6667
20	0	1	.0667	.0333	.0333
		1.5	.2667	.2333	.2667
		2	.4667	.5000	.6000
		3	.7333	.8666	.9000
		4	.9333	.9333	1.0000
		5	.9000	.9333	1.0000
	20	1	.0000	.0000	.0333
		1.5	.0667	.1333	.2000
		2	.3000	.3000	.4333
		3	.8000	.9000	.9333
		4	.8667	.9667	1.0000
		5	.9333	.9667	.9667
	40	1	.0333	.0000	.0000
		1.5	.1667	.2000	.2333
		2	.2000	.2333	.4000
		3	.5667	.6333	.7667
		4	.8333	.8667	.9333
		5	.9333	.9667	1.0000

APPENDIX E

Results for a test of normality for four response functions.

Given is the observed significance level of a test of the null hypothesis that the 30 data points for each combination of factors are a random sample taken from a normal distribution. The four response functions are:

$$\begin{aligned} P &= \text{integer p-value} \\ LP &= \log(P) \\ LLP &= \log[\log(10*P)] \\ ASP &= \text{acrsin}(\sqrt{p}) \end{aligned}$$

The levels of sample size (S), censoring (C), and mean for the second survival distribution (M) are also reported. The 30 p-values for each treatment combination were those yielded by Gehan's test. This was done out of a cost consideration since the p-values for the other tests followed roughly the same trends. In addition, the variance of the thirty values for the second response function are reported as Var(LP). Significance levels not reported were less than .01.

S	C	M	P	LP	LLP	ASP	Var(LP)
10	0	1	.193	--	--	.618	0.908
		1.5	.018	--	--	.380	2.012
		2	--	.111	.130	--	1.853
		3	--	.927	.436	--	3.246
		4	--	.562	.649	--	3.094
		5	--	.122	.441	--	3.584
	20	1	.032	--	--	.066	0.550
		1.5	--	.223	.152	--	0.859
		2	--	.142	.066	--	2.439
		3	--	.491	.066	--	2.020
		4	--	.313	--	--	3.331
		5	--	.299	.013	--	3.465
	40	1	.047	--	--	.240	0.467
		1.5	.075	--	--	.550	1.444
		2	--	.107	.061	--	1.350
		3	--	.019	--	.061	1.876
		4	--	.439	.057	--	3.111
		5	--	.041	.010	--	3.165
20	0	1	.267	--	--	.622	1.022
		1.5	--	--	--	.039	1.907
		2	--	.085	--	--	3.489
		3	--	--	.059	--	4.356
		4	--	--	.357	--	4.310
		5	--	--	--	--	5.963
	20	1	.091	--	--	.504	.0506
		1.5	--	.050	.037	.037	1.045
		2	--	--	--	.012	4.247
		3	--	.135	.079	--	2.498
		4	--	.031	.041	--	4.282
	40	1	.190	--	--	.626	0.669
		1.5	--	--	--	.027	2.216
		2	--	.056	--	.123	1.896
		3	--	.043	--	--	7.077
		4	--	.477	.296	--	5.400
		5	--	.823	.047	--	3.311

APPENDIX F

Mean p-values of each test for each combination of factors.

The factors are sample size (S), the level of censoring (C), and the mean of the second survival distribution (M). Each mean is an average of thirty p-values.

S	C	M	Gehan's	Logrank	Likelihood Ratio
10	0	1	.5086	.5092	.5097
		1.5	.3424	.2908	.2679
		2	.2147	.2014	.1866
		3	.1221	.1150	.0797
		4	.0543	.0302	.0168
		5	.0368	.0231	.0116
	20	1	.5301	.5443	.5202
		1.5	.3689	.3358	.3434
		2	.3037	.2756	.2300
		3	.1529	.1286	.12066
		4	.0730	.0589	.0411
		5	.0565	.0447	.0209
	40	1	.5373	.5667	.5568
		1.5	.3775	.4514	.4164
		2	.3424	.3096	.2270
		3	.2172	.2067	.1785
		4	.1776	.1368	.0817
20	0	1	.4738	.5084	.4499
		1.5	.2724	.2865	.2926
		2	.1687	.1526	.1362
		3	.0395	.0160	.0101
		4	.0172	.0046	.0008
		5	.0238	.0094	.0013
	20	1	.5849	.6337	.6567
		1.5	.3661	.3487	.3518
		2	.2798	.2049	.1985
		3	.0339	.0204	.0131
		4	.0188	.0092	.0044
		5	.0082	.0074	.0079
	40	1	.5446	.5277	.5277
		1.5	.4738	.3820	.3248
		2	.2460	.2390	.2144
		3	.1169	.1333	.1068
		4	.0477	.0510	.0213
		5	.0156	.0111	.0020

APPENDIX G

Mean differences for the comparison of test procedures and the observed significance level for a t-test that the mean difference is zero.

Given below are the means for the following differences:

$$\text{Diff1} = \log(p_1) - \log(p_2)$$

$$\text{Diff2} = \log(p_1) - \log(p_3)$$

$$\text{Diff3} = \log(p_2) - \log(p_3)$$

where p_1, p_2, and p_3 are the integer p-values obtained from Gehan's test, the logrank test, and the likelihood ratio test, respectively. These means are reported for each of the factors of sample size (S), level of censoring (C), and the second distributional mean (M) as well as the combination of these three factors. Along with the mean differences is given the observed significance level for a t-test of the hypothesis that this mean difference is zero.

S	DIFF1	PROB > [T[
	LSMEAN	H0:LSMEAN=0
10	0.14747891	0.0001
20	0.34182701	0.0001

S	DIFF2	PROB > [T[
	LSMEAN	H0:LSMEAN=0
10	0.83446067	0.0001
20	1.03493855	0.0001

```
 S          DIFF3    PROB > [T[
           LSMEAN   H0:LSMEAN=0

10     0.68698176     0.0001
20     0.69311153     0.0001


 C          DIFF1    PROB > [T[
           LSMEAN   H0:LSMEAN=0

0      0.30091683     0.0001
20     0.25264215     0.0001
40     0.18039990     0.0001


 C          DIFF2    PROB > [T[
           LSMEAN   H0:LSMEAN=0

0      0.95792037     0.0001
20     0.93774790     0.0001
40     0.90843056     0.0001


 C          DIFF3    PROB > [T[
           LSMEAN   H0:LSMEAN=0

0      0.65700354     0.0001
20     0.68510575     0.0001
40     0.72803065     0.0001


  M          DIFF1    PROB > [T[
            LSMEAN   H0:LSMEAN=0

1      -0.05653454     0.2344
2       0.15177354     0.0014
3       0.37910563     0.0001
4       0.46810607     0.0001
5       0.43492668     0.0001
1.5     0.09054038     0.0570

  M          DIFF2    PROB > [T[
            LSMEAN   H0:LSMEAN=0

1      -0.01964329     0.8477
2       0.67267922     0.0001
3       1.25314812     0.0001
4       1.68265883     0.0001
5       1.67385875     0.0001
1.5     0.34549602     0.0008
```

```
         M            DIFF3     PROB > [T[
                     LSMEAN     H0:LSMEAN=0

        1          0.03689124     0.6620
        2          0.52090568     0.0001
        3          0.87404249     0.0001
        4          1.21455276     0.0001
        5          1.23893207     0.0001
        1.5        0.25495564     0.0026


  S      C      M            DIFF1       PROB > [T[
                            LSMEAN      H0:LSMEAN=0

 10      0      1         0.00483550       0.9669
 10      0      2         0.08376938       0.4719
 10      0      3         0.27192433       0.0197
 10      0      4         0.36535305       0.0017
 10      0      5         0.39505063       0.0007
 10      0      1.5       0.13644157       0.2414
 10     20      1        -0.04470125       0.7010
 10     20      2         0.04071987       0.7265
 10     20      3         0.22372159       0.0549
 10     20      4         0.22748344       0.0509
 10     20      5         0.24314856       0.0370
 10     20      1.5       0.14519652       0.2125
 10     40      1        -0.12116607       0.2982
 10     40      2         0.07414776       0.5243
 10     40      3         0.20864465       0.0734
 10     40      4         0.25402865       0.0293
 10     40      5         0.35777812       0.0022
 10     40      1.5      -0.21175593       0.0692
 20      0      1        -0.04550831       0.6959
 20      0      2         0.26770824       0.0217
 20      0      3         0.69705709       0.0001
 20      0      4         0.68683874       0.0001
 20      0      5         0.76400358       0.0001
 20      0      1.5      -0.01647179       0.8875
 20     20      1        -0.12789645       0.2721
 20     20      2         0.32340365       0.0056
 20     20      3         0.55142815       0.0001
 20     20      4         0.80576604       0.0001
 20     20      5         0.50582901       0.0001
 20     20      1.5       0.13760664       0.2374
 20     40      1        -0.00477065       0.9673
 20     40      2         0.12089235       0.2992
 20     40      3         0.32185798       0.0058
 20     40      4         0.46916651       0.0001
 20     40      5         0.34375020       0.0032
 20     40      1.5       0.35222528       0.0025
 S     C      M           DIFF2       PROB > [T[
                         LSMEAN      HO:LSMEAN=0

10     0     1        0.04206661       0.8667
10     0     2        0.66248290       0.0083
10     0     3        0.96364721       0.0001
10     0     4        1.68912455       0.0001
10     0     5        1.94398621       0.0001
10     0     1.5      0.39180115       0.1181
10    20     1       -0.00161149       0.9949
10    20     2        0.63579543       0.0113
10    20     3        0.94711555       0.0002
10    20     4        1.66398644       0.0001
10    20     5        1.49584269       0.0001
10    20     1.5      0.34287811       0.1713
10    40     1       -0.05189432       0.8359
10    40     2        0.71223533       0.0046
10    40     3        0.86479493       0.0006
10    40     4        1.39396111       0.0001
10    40     5        1.23174880       0.0001
10    40     1.5      0.09233083       0.7125
20     0     1        0.06199837       0.8046
20     0     2        0.89172307       0.0004
20     0     3        1.74140413       0.0001
20     0     4        1.62531304       0.0001
20     0     5        1.51032077       0.0001
20     0     1.5     -0.02882355       0.9084
20    20     1       -0.12720818       0.6117
20    20     2        0.52185526       0.0375
20    20     3        1.85048562       0.0001
20    20     4        2.07110296       0.0001
20    20     5        1.44847585       0.0001
20    20     1.5      0.40425653       0.1069
20    40     1       -0.04121075       0.8694
20    40     2        0.61198332       0.0147
20    40     3        1.15144129       0.0001
20    40     4        1.65246489       0.0001
20    40     5        2.41277817       0.0001
20    40     1.5      0.87053308       0.0005
```

```
 S    C     M          DIFF3     PROB > [T[
                      LSMEAN    H0:LSMEAN=0

10    0    1      0.03723111      0.8571
10    0    2      0.57871352      0.0052
10    0    3      0.69172289      0.0008
10    0    4      1.32377151      0.0001
10    0    5      1.54893558      0.0001
10    0    1.5    0.25535959      0.2168
10    20   1      0.04308976      0.8349
10    20   2      0.59507556      0.0041
10    20   3      0.72339396      0.0005
10    20   4      1.43650300      0.0001
10    20   5      1.25269413      0.0001
10    20   1.5    0.19768159      0.3390
10    40   1      0.06927175      0.7375
10    40   2      0.63808757      0.0021
10    40   3      0.65615029      0.0015
10    40   4      1.13993246      0.0001
10    40   5      0.87397068      0.0001
10    40   1.5    0.30408676      0.1414
20    0    1      0.10750668      0.6030
20    0    2      0.62401483      0.0026
20    0    3      1.04434704      0.0001
20    0    4      0.93847430      0.0001
20    0    5      0.74631719      0.0003
20    0    1.5   -0.01235177      0.9523
20    20   1      0.00068827      0.9973
20    20   2      0.19845161      0.3371
20    20   3      1.29905747      0.0001
20    20   4      1.26533693      0.0001
20    20   5      0.94264684      0.0001
20    20   1.5    0.26664989      0.1972
20    40   1     -0.03644011      0.8601
20    40   2      0.49109097      0.0177
20    40   3      0.82958331      0.0001
20    40   4      1.18329837      0.0001
20    40   5      2.06902797      0.0001
20    40   1.5    0.51830780      0.0123
```

AN INVESTIGATION INTO THE POWER OF THREE TESTS USED TO COMPARE SURVIVAL DISTRIBUTIONS

by

JOHN STEVEN GATSCHET

B.A., Saint Louis University, 1985

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fullfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1987

ABSTRACT

This study compares the power of Gehan's generalized Wilcoxon test, the logrank test, and the likelihood ratio test and develops models for the power function of each. The factors varied in this study are sample size, the level of censoring, and the ratio of the two survival distribution means. The underlying survival distributions are exponential. The type of censoring employed is random censoring. The samples are generated by computer simulation using 30 iterations for each of the 36 factor combinations. The computer cost to perform iterations of 400 or 1000 would be prohibitive. This presented the opportunity to apply principles of analysis of experimental data to simulation studies.

Three separate analyses are performed: analysis of power values, splitplot analysis of the transformed p-values, and analysis of the difference between transformed p-values. The design used is completely randomized. Each analysis suggests that the likelihood ratio test is the most powerful. The logrank test is the second most powerful. Gehan's test is the least powerful. Sample size and the level of censoring do not affect these findings.

The power functions were developed using a method similar to backward stepwise regression. These functions appear relatively accurate in estimating the true power functions. A large number of iterations for each factor combination would improve the accuracy of these functions yet also greatly increase computer costs.