A Monte Carlo Study of Some Two-Sample Rank Tests With Censored Data

ROBERT B. LATTA*

The powers of several nonparametric tests for the two-sample problem with censored data are compared by simulation. The tests studied include Gehan's, Efron's, and Peto's, and Prentice's generalized Wilcoxon tests, along with the logrank test. The test with the greatest power changes with the sample sizes, censoring mechanism, and distribution of the random variables. The test that performed the best overall was the Peto-Prentice generalized Wilcoxon statistic with an asymptotic variance estimate.

KEY WORDS: Censored data; Gehan's test; Logrank test; Nonparametric tests; Wilcoxon test.

1. INTRODUCTION

In survival studies an investigator frequently wants to determine if subjects from one population live longer than those from a second population. This problem can be formalized in the following way. Let F denote the survival function of the lifetimes of the subjects from the first population and let G denote that for the second population. Note that $F = 1 -$ cumulative distribution function of the lifetime. Then the null hypothesis becomes H_0: $F(x) = G(x)$ and the one-sided alternative becomes H_A: $F(x) \geq G(x)$ for all x with strict inequality for some x. The problem is further complicated when the lifetimes are right censored.

Several generalizations of the Wilcoxon statistic for the two-sample problem with censored data have been proposed. The generalization that is presently most commonly used is the Gehan (1965) test with a variance estimate based on the permutation distribution. Breslow (1970) proposed an asymptotic variance estimate for Gehan's statistic. Prentice and Marek (1979) proposed an additional variance estimate that is similar to Breslow's estimate.

Efron (1967) raised several objections to the Gehan test and proposed an alternative generalization. Peto and Peto (1972) proposed a generalized Wilcoxon statistic that uses generalized ranks or scores computed from the Kaplan and Meier (1958) estimate of the survival function. The variance estimate they proposed is based on the permutation distribution. Prentice (1978) proposed a similar test, along with an asymptotic estimate for the variance. An additional estimate for the variance was also proposed by Prentice and Marek (1979). This variance estimate is similar to the one they proposed for Gehan's test.

A generalization of Savage's (1956) test is the most commonly used test for the two-sample problem with censored data. Mantel (1966) and Cox (1972) proposed the same generalization, the logrank statistic, and the same variance estimate. Prentice (1978) discussed linear rank tests with censored data in general and illustrated the similarities between the generalized Wilcoxon tests and the logrank test. Both Prentice (1978) and Peto and Peto (1972) noted that an estimate for the variance of the logrank statistic could also be obtained from the permutation distribution. Prentice also gave an asymptotic variance estimate that is the same as Cox's and Mantel's when there are no ties among the uncensored observations.

Lee, Desu, and Gehan (1975) performed a simulation study that compares the power of Gehan's test with the permutation variance estimate, Peto and Peto's test with the permutation variance estimate, and several other tests. They restricted their study to the case in which both samples experienced the same censoring mechanism and both sample sizes were the same.

This paper compares the power of all of the above-mentioned tests. Various censoring mechanisms and sample sizes were used. The simulation studies include cases in which the samples experience different censoring mechanisms and cases in which the sample sizes are not equal.

2. NOTATION AND DESCRIPTION OF TESTS

2.1 Gehan's Test, Peto's Test, and the Logrank Test

To describe these tests, this paper uses notation similar to the notation that Prentice used in his two articles. Let $t_1 < t_2 < \cdots < t_k$ denote the ordered distinct uncensored survival times for the combined samples. Let d_{1i} and d_{2i} for $i = 1, 2, \ldots, k$ be the number of population 1 and population 2 deaths, respectively, at t_i. Set $d_i = d_{1i} + d_{2i}$ for $i = 1, 2, \ldots, k$. Set $d_0 = d_{1,0} = d_{2,0} = 0$. Let e_{1i} and e_{2i} for $i = 0, 1, 2, \ldots, k$ be, respectively, the number of population 1 and population 2 right-censored observations that fall in the interval $[t_i, t_{i+1})$ with $t_0 = -\infty$ and $t_{k+1} = +\infty$. Set $e_i = e_{1i} + e_{2i}$. Let n_{1i} and n_{2i} be the number of population 1 and population 2 subjects with survival times known to be greater than or equal to

[*] Robert B. Latta is Mathematical Statistician, Office of the Consumption Data System, Energy Information Administration, U.S. Department of Energy, Washington, DC 20585.

© Journal of the American Statistical Association
September 1981, Volume 76, Number 375
Theory and Methods Section

t_i. Set $n_i = n_{1i} + n_{2i}$. Let m_1 and m_2 be the total number of population 1 and population 2 subjects, respectively. Set $N = m_1 + m_2$.

Consider statistics of the form

$$v = \sum_{i=0}^{k} (d_{2i}c_i + e_{2i}C_i). \qquad (2.1)$$

First, assume that $d_i \equiv 1$ or there are no ties among the uncensored observations. If $c_i = n_i - i$ and $C_i = -i$, $i = 1, 2, \ldots, k$, then v is equal to Gehan's generalized Wilcoxon statistic. The logrank test is given by the scores $c_i = 1 - \sum_{j=1}^{i} n_j^{-1}$ and $C_i = -\sum_{j=1}^{i} n_j^{-1}$, $i = 1, 2, \ldots, k$. Peto's generalized Wilcoxon statistic and Prentice's generalized Wilcoxon statistic both use scores of the form $c_i = 2F_i - 1$ and $C_i = F_i - 1$, $i = 1, 2, \ldots, k$, where F_i is an estimate of the survival function at time t_i. Let $\hat{F}_i = \Pi_{j=1}^{i} (n_j - 1)/n_j$, $i = 1, \ldots, k$. Set $\hat{F}_0 = 1.0$. Then Peto and Peto used $F_i = (\hat{F}_i + \hat{F}_{i-1})/2$. Let $\tilde{F}_i = \Pi_{j=1}^{i} [n_j/(n_j + 1)]$, $i = 1, 2, \ldots, k$. Then Prentice used $F_i = \tilde{F}_i$. It is obvious that these two sets of scores will be very close to each other. From now on in this paper, the scores with $F_i = \tilde{F}_i$ will be used and the resulting statistic will be called the Peto-Prentice generalized Wilcoxon statistic.

When ties among the uncensored observations are present, the scores will first be calculated as if there are no ties, then the average score will be used for the tied observations.

Prentice noted that the expected value of v is zero for all three sets of scores. Three methods of obtaining a variance estimate for v will be discussed.

The simplest way to obtain a variance estimate is to assume that the censoring mechanisms are the same for both populations. Hence each of the $\binom{N}{m_1}$ divisions of the subjects into two populations of sizes m_1 and m_2 is possible and equally likely. The permutation distribution can then be used to obtain the variance of v. This yields $V = m_1m_2[N(N - 1)]^{-1} \sum_{i=0}^{k} (d_ic_i^2 + e_iC_i^2)$ as a variance estimate for v. This estimate will be called the permutation variance estimate.

If the censoring mechanisms for the two samples are not the same, then each of the divisions may not be equally likely. Hence the previous permutation argument is not valid in this case. Prentice and Marek (1979) used a conditional permutation approach. They expressed the statistic in the form

$$v = \sum_{j=1}^{k} w_i(d_{2i} - d_in_{2i}n_i^{-1}). \qquad (2.2)$$

The Gehan scores correspond to $w_i = n_i$, the logrank scores to $w_i = 1$, and the Peto-Prentice scores to $w_i = \tilde{F}_i$.

Prentice and Marek noted that the terms in (2.2) are uncorrelated and hence the variance of v can be obtained by summing the variances of the individual terms. The variance for each term can be obtained by using the partial likelihood approach of Cox (1972). This approach uses the permutation distribution of d_{2i} where d_i, n_{1i}, and n_{2i} are held fixed and all subjects whose lifetimes are known to be at least t_i are considered to be equally likely to have died at t_i. The hypergeometric distribution gives the null hypothesis conditional variance of the terms. This yields

$$\begin{aligned} V = \sum_{i=1}^{k} & d_iw_i^2(n_{2i}/n_i) \\ & \times [(n_i - n_{2i})/n_i][(n_i - d_i)/(n_i - 1)], \end{aligned} \qquad (2.3)$$

where the final term in the sum is defined to take the value zero if $n_k = 1$. This estimate will be called the conditional permutation variance estimate.

Prentice (1978) also derived asymptotic variance estimates for the logrank and Peto-Prentice statistics based on the second derivative of the log-likelihood from the marginal distribution of the rank vector. The estimate for the variance of the logrank statistic will be the same as the conditional permutation variance estimate when there are no ties among the uncensored observations. A variance estimate that is similar in nature can also be given for the Gehan statistic. It too will be the same as the conditional permutation variance estimate when there are no ties among the uncensored observations.

The asymptotic estimate for the Peto-Prentice generalized Wilcoxon statistic will in general be different from the conditional permutation estimate, even if there are no ties. When there are no ties among the uncensored observations the asymptotic estimate is given by

$$\begin{aligned} V = \sum_{i=1}^{k} \Bigg\{ & \tilde{F}_i(1 - a_i)X_i - (a_i - \tilde{F}_i)X_i \\ & \times \left(\tilde{F}_iX_i + 2\sum_{j=i+1}^{k} \tilde{F}_jX_j\right)\Bigg\}, \end{aligned}$$

where $a_i = \Pi_{j=1}^{i} (n_j + 1)/(n_j + 2)$ and $X_i = 2d_{2i} + e_{2i}$. When there are ties, Prentice suggested replacing $\tilde{F}_i$ and a_i by the average score in the variance estimate. The average scores will be the same ones used in calculating the statistic.

2.2 Efron's Generalized Wilcoxon Test

An algorithm was developed for computing Efron's statistic and an estimate of its variance. A modification of his statistic that allows the largest observation in each group to be censored was also computed. In the simulation studies described below, neither the original Efron statistic nor the modification handled heavy late censoring adequately. Also, the original Efron statistic is adversely affected by either unequal sample sizes or different censoring mechanisms. Even with equal sample sizes, equivalent censoring mechanisms, and only light-to-moderate late censoring, the variance estimates for the two statistics did not appear to be adequate for small or moderate sample sizes.

The Efron statistic, the modification, and the variance estimates require that the tails of the distributions of the lifetimes for both groups be estimated. When either of the tail estimates is inadequate, one or more of the previously mentioned problems can occur. In particular, truncation of the lifetimes at some level above which a considerable number of the lifetimes normally fall will adversely affect either of the two Efron statistics.

The two Efron statistics were included in the simulation study, but because of their overall poor performance the results were not included in the final summaries.

3. MONTE CARLO SAMPLING PROCEDURE

The performance of the tests was studied using computer simulation. The SRAND procedure of the IMSL package was used on an IBM 370 computer at the University of Kentucky Computer Center to generate the random numbers used in the study. Four independent sets of independent random variables—$(X_1{}^o, X_2{}^o, \ldots, X_{m_1}{}^o)$, $(Y_1{}^o, Y_2{}^o, \ldots, Y_{m_2}{}^o)$, $(U_1, U_2, \ldots, U_{m_1})$, and $(V_1, V_2, \ldots, V_{m_2})$—were generated for each case studied. The censoring was simulated by comparing $X_i{}^o$ with U_i and $Y_j{}^o$ with V_j. If $X_i{}^o < U_i$, then X_i is set equal to $X_i{}^o$ and the observation is treated as if it were uncensored. If $X_i{}^o > U_i$, then X_i is set equal to U_i and the observation is treated as if it were censored at U_i. An analogous procedure was used to determine Y_j. The resulting X_i's and Y_j's are the observed lifetimes for group one and group two, respectively.

The properties of the various tests were compared by using lognormal, exponential, and Weibull distributions for the $X_i{}^o$'s and $Y_j{}^o$'s. For the lognormal distribution, the $Y_j{}^o$'s were always distributed as lognormal random variables with parameters $\mu = 0$ and $\sigma^2 = 1$. The $X_i{}^o$'s were distributed as lognormal random variables with $\sigma^2 = 1$, but μ was varied. For the Weibull distribution, the Y_j's were always distributed as Weibull random variables with shape parameter $\gamma = 4.0$ and scale parameter $\lambda = 1.0$. The $X_i{}^o$'s were distributed as Weibull random variables with $\gamma = 4.0$, but λ were varied. For exponential distribution, the $Y_j{}^o$'s were always distributed as exponential random variables with scale parameter $\lambda = 1.0$. The $X_i{}^o$'s were distributed as exponential random variables with various values of λ.

Various censoring mechanisms were simulated by using various distributions for the U_i's and V_j's. The censoring pattern resulting from a clinical trial in which the patient input rate is constant over time was simulated by allowing U_i and V_j to be uniformly distributed over $(0, T)$. If T equals 1.5936 and $X_i{}^o$ and $Y_j{}^o$ are exponential random variables with $\lambda = 1.0$, then the probability that a given observation is censored equals .5. If the expected value of $X_i{}^o$ or $Y_j{}^o$ is decreased or increased, then the probability that an observation is censored is decreased or increased.

Sample sizes of $m_1 = 10$, $m_2 = 10$; $m_1 = 50$, $m_2 = 50$; $m_1 = 50$, $m_2 = 10$; and $m_1 = 10$, $m_2 = 50$ were used. One thousand repetitions were used in each instance. For each repetition, the values of the statistics were calculated from the same simulated sample. For each statistic, the number of repetitions in which the normalized form of the statistic exceeded 1.282, 1.645, 2.326, 3.090, and other critical points were recorded. These correspond to significance levels of .1, .05, .01, and .001, respectively. The same procedure was followed for the number of repetitions in which the normalized form of the statistic was less than -1.282, -1.645, -2.326, and -3.090. To save space, only the results for 1.645 and -1.645 are reported for the power studies.

The test statistics and variance estimates used in the Monte Carlo simulations were Gehan's statistic with the permutation variance estimate, Gehan's statistic with the conditional permutation variance estimate, Peto-Prentice statistic with the permutation variance estimate, Peto-Prentice statistic with the conditional permutation variance estimate, Peto-Prentice statistic with the asymptotic variance estimate, logrank statistic with the permutation variance estimate, and the logrank statistic with the conditional permutation variance estimate.

4. POWER STUDY RESULTS

4.1 Case 1: Uniform Censoring in Both Samples

In case 1, both U_i and V_j are always distributed as uniform (0, 1.5936) random variables. The distribution of V_j does not change when the distribution of Y_j is changed. Table 1 summarizes the results for case 1. The columns headed by LT give the number of normalized statistics that are less than -1.645. The columns headed by UT give the number of normalized statistics that are greater than 1.645. Both correspond to levels of significance of .05. In all simulation studies, positive values for the test statistic imply that the first sample tends to be larger than the second sample. When $X_i{}^o$ and $Y_j{}^o$ are identically distributed, both of these numbers ideally should be close to 50. All the results given in the same row were obtained from the same set of simulations. The results in different rows correspond to different simulation runs.

4.2 Case 2: No Censoring in First Sample, Uniform Censoring in Second Sample

In case 2, U_i is always infinity. In the second sample, V_j is always distributed as a uniform (0, 1.5936) random variable. For this case some of the assumptions used in deriving the permutation distribution for the null hypothesis do not hold. Table 2 summarizes the results for case 2. The format of Table 2 is the same as that in Table 1.

5. ADDITIONAL STUDIES OF THE NULL DISTRIBUTION

If there is no censoring, the Gehan statistic and the Peto-Prentice statistic will be equal to the Wilcoxon statistic. The permutation variance estimates for the Gehan and Peto-Prentice statistics will always equal the variance of the Wilcoxon statistic. The conditional permutation variance estimates and the asymptotic variance estimates

Table 1. Case 1: Uniform Censoring in Both Samples

Type of Distribution	m_1	m_2	λ or μ First Sample	λ or μ Second Sample	Gehan Permutation Variance LT	UT	Gehan Conditional Permutation Variance LT	UT	Peto-Prentice Permutation Variance LT	UT	Peto-Prentice Conditional Permutation Variance LT	UT	Peto-Prentice Asymptotic Variance LT	UT	Logrank Permutation Variance LT	UT	Logrank Conditional Permutation Variance LT	UT
Exponential	10	10	1.0	1.0	54	57	53	56	58	50	59	51	59	51	55	44	61	52
			4.0	1.0	0	466	0	465	0	491	0	490	0	492	0	501	0	501
			.5	1.0	262	5	265	6	276	4	280	5	280	5	294	4	304	4
	50	10	1.0	1.0	43	54	33	67	46	51	32	71	50	54	50	42	37	66
			2.5	1.0	0	463	0	486	0	489	0	523	0	487	0	492	0	536
			.625	1.0	224	8	188	9	243	7	209	9	256	8	281	6	229	9
	50	50	1.0	1.0	44	56	44	57	44	62	44	62	44	62	48	61	50	63
			1.6	1.0	3	407	3	407	1	459	1	459	1	458	1	484	1	484
			.625	1.0	460	0	462	0	488	0	491	0	490	0	508	1	511	1
Weibull γ = 4.0	10	10	1.0	1.0	62	42	61	41	58	39	61	40	60	43	56	37	67	51
			4.0	1.0	2	380	2	388	2	420	2	423	2	431	2	436	2	451
			.25	1.0	523	1	532	1	550	0	565	0	556	0	558	0	607	0
	50	10	1.0	1.0	35	65	27	86	37	59	24	83	42	64	52	41	28	100
			2.5	1.0	1	369	1	401	1	399	1	449	2	399	1	369	0	495
			.4	1.0	394	0	354	1	444	0	393	0	496	0	557	0	473	0
	50	50	1.0	1.0	41	46	42	47	43	48	45	49	44	48	48	54	51	54
			1.6	1.0	3	339	3	341	1	368	1	372	1	371	0	412	0	417
			.625	1.0	374	1	376	1	415	1	413	1	415	1	470	1	476	1
Log-normal	10	10	.0	.0	52	53	54	57	45	56	46	58	47	56	40	57	44	54
			1.0	.0	0	408	1	411	0	419	0	419	0	416	0	423	0	425
			−1.0	.0	574	0	577	0	562	0	571	0	565	0	542	0	552	0
	50	10	.0	.0	54	49	43	57	60	54	44	63	60	51	58	49	46	67
			.625	.0	0	409	0	434	0	389	0	419	0	400	0	367	0	413
			− .625	.0	391	1	354	1	427	1	379	1	446	1	430	1	372	1
	50	50	.0	.0	60	42	60	42	67	48	66	48	66	48	57	56	57	56
			.4	.0	1	445	1	442	2	440	2	442	2	441	4	426	4	427
			− .4	.0	503	1	503	1	507	1	509	1	510	1	497	0	500	0

Table 2. Case 2: No Censoring in First Sample, Uniform Censoring in Second Sample

Type of Distribution	m_1	m_2	γ or μ First Sample	γ or μ Second Sample	Gehan Permutation Variance LT	UT	Gehan Conditional Permutation Variance LT	UT	Peto-Prentice Permutation Variance LT	UT	Peto-Prentice Conditional Permutation Variance LT	UT	Peto-Prentice Asymptotic Variance LT	UT	Logrank Permutation Variance LT	UT	Logrank Conditional Permutation Variance LT	UT
Exponential	10	10	1.0	1.0	49	46	49	53	49	42	47	56	50	55	46	22	46	62
			4.0	1.0	0	485	0	538	0	466	0	563	0	550	0	293	0	594
			.5	1.0	298	2	293	2	305	2	303	2	306	2	352	1	342	2
	10	50	1.0	1.0	75	68	72	49	71	75	78	50	56	69	94	78	83	45
			1.6	1.0	4	249	5	200	4	286	4	218	2	273	6	321	5	244
			.625	1.0	308	5	313	3	334	5	335	2	292	3	369	6	370	5
	50	50	1.0	1.0	48	51	48	55	49	48	52	55	54	54	45	28	53	54
			1.6	1.0	0	447	0	468	0	453	0	481	0	477	0	339	0	502
			.625	1.0	551	0	551	0	569	0	570	0	573	0	596	0	612	0
Weibull γ = 4.0	10	10	1.0	1.0	57	39	57	57	54	40	57	58	68	54	65	27	62	73
			4.0	1.0	0	489	1	611	0	485	0	628	0	591	0	318	0	669
			.4	1.0	360	3	361	3	371	3	370	4	417	3	461	0	447	4
	10	50	1.0	1.0	93	63	68	28	93	66	65	29	55	39	103	82	76	40
			2.0	1.0	4	475	0	351	1	513	0	383	0	429	3	602	2	447
			.5	1.0	525	1	444	0	549	1	482	0	434	0	617	4	550	2
	50	50	1.0	1.0	46	42	51	54	48	42	52	54	55	51	48	32	54	69
			1.6	1.0	0	404	0	456	0	414	0	472	0	451	0	384	0	528
			.625	1.0	455	0	466	0	474	0	487	0	508	0	548	0	563	0
Log-normal	10	10	.0	.0	46	40	48	55	49	38	50	55	52	51	45	24	52	59
			.8	.0	2	310	2	386	0	271	2	395	2	379	0	125	1	401
			− .8	.0	437	1	435	2	444	0	438	2	448	2	418	0	414	2
	10	50	.0	.0	76	62	69	40	82	71	72	42	60	54	86	72	67	48
			.625	.0	1	464	1	377	1	476	1	396	1	427	1	448	1	379
			− .625	.0	555	0	544	0	560	1	544	0	510	1	528	1	496	0
	50	50	.0	.0	49	47	50	40	47	43	51	56	53	53	30	18	46	53
			.4	.0	1	448	1	497	1	425	1	495	1	491	1	231	1	482
			− .4	.0	511	0	516	0	507	0	511	0	514	0	414	0	478	0

Table 3. Simulations With No Censoring

			(.001)	(.01)	(.05)	(.10)	(.10)	(.05)	(.01)	(.001)
m_1	m_2	Test Statistic	<-3.090	<-2.326	<-1.645	<-1.282	>1.282	>1.645	>2.326	>3.090
10	10	A: Gehan or Peto-Prentice with permutation variance	6	92	539	1,122	1,101	545	71	2
		B: Gehan or Peto-Prentice with conditional permutation variance	7	111	539	1,122	1,101	546	98	8
		C: Peto-Prentice with asymptotic variance	7	96	540	1,072	1,044	548	84	4
		D: Logrank with permutation variance	0	45	523	1,095	1,136	515	57	0
		E: Logrank with conditional permutation variance	18	140	623	1,142	1,195	616	142	17
50	10	A	5	78	520	1,020	1,052	526	72	4
		B	1	39	387	889	1,207	668	185	29
		C	16	108	541	1,039	1,077	540	98	7
		D	32	158	612	1,079	940	335	7	0
		E	8	67	391	843	1,362	769	222	33
50	50	A	12	84	495	1,033	1,012	505	89	10
		B	12	89	497	1,033	1,012	506	92	10
		C	12	87	497	1,032	1,007	506	92	10
		D	9	89	490	1,002	1,042	524	87	7
		E	14	112	513	1,010	1,063	546	108	10

Table 4. Simulations With Heavy Early Censoring

			(.001)	(.01)	(.05)	(.10)	(.10)	(.05)	(.01)	(.001)
m_1	m_2	Test Statistic	<-3.090	<-2.326	<-1.645	<-1.282	>1.282	>1.645	>2.326	>3.090
50	50	A: Gehan's with permutation variance	0	0	1	60	58	0	0	0
		B: Gehan's with conditional permutation variance	0	0	1	49	47	0	0	0
		C: Peto-Prentice with permutation variance	1	12	66	104	107	62	14	0
		D: Peto-Prentice with conditional permutation variance	1	12	55	106	107	51	14	0
		E: Peto-Prentice with asymptotic variance	1	12	56	106	107	56	14	0
		F: Logrank with permutation variance	1	9	57	115	107	52	8	0
		G: Logrank with conditional permutation variance	2	14	62	119	103	59	13	0

will not in general equal the Wilcoxon variance, but the conditional permutation variance estimates for the Gehan and Peto-Prentice statistics will be equal to each other.

Table 3 summarizes the results of a Monte Carlo study in which there is no censoring. For this study, both X_i^o and Y_j^o are always uniform (0, 1) random variables and U_i and V_j are always equal to infinity. The format of Table 3 differs from the format of previous tables. The columns of Table 3 give the number of simulations in which the standardized statistic is greater than positive critical values or less than negative critical values. Positive values of the statistic imply that the observations in the first sample tend to be larger than the observations in the second sample. The critical values used were -3.090, -2.326, -1.645, -1.282, 1.282, 1.645, 2.326, and 3.090. These correspond to levels of significance of .001, .01, .05, .1, .1, .05, .01, .001, respectively. Ten thousand repetitions were used for this study. Hence the ideal number of significant statistics for the critical values listed above would be 10, 100, 500, 1,000, 1,000, 500, 100, 10, respectively. All the results given for the same sample size were obtained from the same set of simulations.

Prentice and Marek (1979) presented an example credited to O'Brien that illustrates one of the disadvantages of the Gehan statistic. In the example, N subjects are followed until the first failure time, after which N' randomly selected subjects are followed until failure. The remaining $N - N' - 1$ are censored at the first failure time. This example illustrates the possible results from heavy early censoring. Table 4 lists the results of a Monte Carlo study of the null distribution that uses this censoring mechanism. For this study $N = 100$ with 50 subjects in each sample and $N' = 12$ with 6 subjects in each sample. The random variables X_i^o and Y_j^o are uniform (0, 1) random variables. The variables U_i and V_j are not used in this censoring scheme. The format of Table 4 is the same as the format of Table 3, except the number of simulations is reduced to 1,000. The ideal number of significant values for each column is one-tenth of the corresponding number in Table 3.

6. CONCLUSIONS

The sample sizes, censoring mechanism, and distribution of the random variables all influenced the relative performance of the test statistics. In addition, the interaction of the factors also influenced the results. The effect of the sample sizes and censoring mechanisms on the null distribution of the test statistic will be discussed first. Then the relative power of the test statistics will be discussed.

The null distribution of the Gehan statistic with either variance estimate appears to be approximately normal for cases 1 and 2 as long as the sample sizes are equal. This trend also shows up in Table 3. On the other hand, Table 4 illustrates that the two Gehan statistics are adversely affected by heavy early censoring. When the sample sizes are unequal, the Gehan statistic with the conditional permutation variance biases the results in favor of the sample with the larger sample size. The effect of unequal sample sizes on the Gehan statistic with the permutation variance is similar when there is heavy censoring, but not as pronounced. In addition, the combination of unequal sample sizes and unequal censoring mechanisms tends to make the Gehan permutation variance estimate too small. This situation is evident in Table 2. When there is no censoring, there does not appear to be any bias due to unequal sample sizes when the permutation variance is used.

The null distribution of the Peto-Prentice statistic with the asymptotic variance appears to be approximately normal for all the simulation studies, even with heavy early censoring or unequal sample sizes. The null distribution of the Peto-Prentice statistic with the permutation variance or the conditional permutation variance is also approximately normal when the sample sizes are equal. With unequal sample sizes the same trends appear using the Peto-Prentice statistic with the permutation variance or the conditional permutation variance as with the Gehan statistic.

The null distribution of the logrank statistic with the conditional permutation variance appears to be approximately normal as long as the sample sizes are equal and the censoring mechanisms are the same, even if there is heavy early censoring. It should be noted that the simulations with no censoring and equal sample sizes imply that the conditional permutation variance estimate may be slightly too small. When the sample sizes are unequal, the statistic is biased toward the larger sample size. When one of the samples experiences censoring and the other does not, then the statistic is biased toward the sample that is not censored or the sample that has the greater number of uncensored observations. Hence these two trends agree with each other.

The bias introduced by the use of the conditional permutation variance estimate when the sample sizes are unequal can be partially explained by examining the case where there is no censoring and no ties. For this case the permutation variance estimate is applicable and depends only on the sample sizes. On the other hand, the conditional permutation variance estimate given by (2.3) varies with the sample. Assume that the sample size for sample 1 is much larger than the sample size for sample 2. Then the conditional permutation variance estimate will be smaller when the population 1 observations tend to be larger than the population 2 observations. Also, the conditional permutation variance estimate will be larger when the opposite occurs. This results in normalized statistics that are too large when the population 1 observations tend to be larger and too small when the opposite occurs.

The same effect occurs with equal sample sizes if one sample is not censored and the other sample experiences censoring that reduces the number of large observations.

Aalen (1978) developed a unified approach to nonparametric tests for the two-sample problem. The Wilcoxon test and the logrank test are special cases of the test he proposed. He also gave an asymptotic variance estimate for the general case. This asymptotic estimate, like the conditional permutation variance estimate, depends on the sample even when there are no ties or censoring. Aalen conjectured that this dependence "means little in practice" p. 719. The results of the simulation study for the conditional permutation variance estimate contradict his conjecture. On the other hand, the asymptotic variance estimate for the Peto-Prentice statistic also depends on the sample. In this case, Aalen's conjecture is supported by the simulation study. As noted previously, Prentice derived both the conditional permutation variance estimate for the logrank test and the asymptotic variance estimate for the Peto-Prentice test by using an asymptotic technique. Hence a general trend cannot be established using this simulation study.

The null distribution of the logrank statistic with the permutation variance appears to be approximately normal as long as the sample sizes are equal and the censoring mechanisms are the same, even if there is heavy early censoring. When the sample sizes are unequal, the statistic is slightly biased toward the sample with the smaller sample size. This trend is most evident in Table 3. Heavy censoring tends to moderate this trend. When one of the samples experiences censoring and the other does not, then the statistic is biased toward the sample that experiences censoring or has the fewer number of uncensored observations. The overall trend for the logrank statistic with the permutation variance is the opposite of the trend of the logrank statistic with the conditional permutation variance. In addition, when only one sample experiences censoring, the permutation variance estimate appears to be too large when the sample sizes are equal and too small when the uncensored sample has a sample size of 10 and the censored sample has a sample size to 50.

If the null distribution of one of the statistics is biased or the variance estimate tends to be too small, then it is hard to compare the power of the tests. But these general trends are evident. For exponential random variables and Weibull random variables with $\gamma = 4.0$, the logrank statistic appears to be the best. For lognormal random variables the Peto-Prentice statistic appears to be the best.

If there is no censoring or light censoring, then Table 3 implies that the Gehan statistic, Peto-Prentice statistic, or logrank statistic, each with any variance estimate, could be used, subject to the provision that unequal sample sizes rule out the use of the conditional permutation variance. The distribution of the random variables could have a big effect on the proper choice of the test statistic.

With heavy censoring and sample sizes that are far apart or censoring mechanisms that differ greatly, the Peto-Prentice statistic with the asymptotic variance should be used. Its null distribution appears to be closer to the normal distribution than the other statistics, and its power tends to be better than or at least close to that for the other statistics for any situation studied.

With heavy censoring, approximately equal sample sizes, and censoring mechanisms that are similar, the choice of the test statistic can be tailored to the random variables involved. The logrank statistic with either variance estimate appears to be best when the random variables are exponential or Weibull. Additional simulations that were not tabled indicate that the relative performance of the test statistics is the same for Weibull random variables with $\gamma = .25$ as for exponential random variables. On the other hand, the Gehan or Peto-Prentice generalized Wilcoxon statistics appear to be the best when the random variables have a lognormal distribution.

The results of the simulation study can be compared to theoretical results about asymptotic relative efficiency. Kalbfleisch and Prentice (1980) presented a summary of the results on efficiency. They noted that "efficiency properties of the censored data rank tests are generally not known" (p. 159). On the other hand, comparison with efficiency properties in the uncensored case is interesting. The locally most powerful rank test when the random variables have an exponential or Weibull distribution is the logrank test. The locally most powerful rank test when the random variables have a lognormal distribution is the normal scores test. The normal scores test is more similar to the Wilcoxon test than the logrank test. Hence the simulation study suggests that the results for the uncensored case carry over into the censored data case.

The test statistic with the overall best performance is the Peto-Prentice statistic with the asymptotic variance estimate. It would even be acceptable to use it when there is no censoring.

[*Received February 1979. Revised March 1981.*]

REFERENCES

AALEN, ODD (1978), "Nonparametric Inference for a Family of Counting Processes," *Annals of Statistics*, 6, 701–726.

BRESLOW, NORMAN E. (1970), "A Generalized Kruskal-Wallis Test for Comparing K Samples Subject to Unequal Censorship," *Biometrika*, 57, 579–594.

COX, DAVID R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

EFRON, BRADLEY (1967), "The Two Sample Problem With Censored Data," *Proceedings of the Fifth Berkeley Symposium*, 4, 831–853.

GEHAN, EDMUND A. (1965), "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples," *Biometrika*, 52, 203–223.

KALBFLEISCH, JOHN D., and PRENTICE, ROSS L. (1980), *The Statistical Analysis of Failure Time Date*, New York: John Wiley.

KAPLAN, E.L., and MEIER, PAUL (1958), "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

LEE, ELISA A., DESU, M.M., and GEHAN, EDMUND A. (1975), "A Monte Carlo Study of the Power of Some Two-Sample Tests," *Biometrika*, 62, 425–432.

MANTEL, NATHAN (1966), "Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration," *Cancer Chemotherapy Reports*, 50, 163–170.

PETO, RICHARD, and PETO, JULIAN (1972), "Asymptotically Efficient Rank Invariant Test Procedures," *Journal of the Royal Statistical Society*, Ser. A, 135, 185–206.

PRENTICE, ROSS L. (1978), "Linear Rank Tests With Right Censored Data," *Biometrika*, 65, 167–179.

PRENTICE, ROSS L., and MAREK, P. (1979), "A Qualitative Discrepancy Between Censored Data Rank Tests," *Biometrics*, 35, 861–867.

SAVAGE, I. RICHARD (1956), "Contributions to the Theory of Rank Order Statistics—The Two Sample Case," *Annals of Mathematical Statistics*, 27, 590–615.