(1965), 52, 1 and 2, p. 203
text-figures
in Great Britain

A generalized Wilcoxon test for comparing arbitrarily singly-censored samples*

By EDMUND A. GEHAN

Birkbeck College, London and the National Institutes of Health, Bethesda

1. Summary

A distribution-free two-sample test is proposed that is an extension of the Wilcoxon test samples with arbitrary censoring on the right. The test is conditional on the pattern observations. The null hypothesis is

$$H_0\colon F_1(t) = F_2(t) \quad (t \leqslant T) \text{ against either}$$
$$H_1\colon F_1(t) < F_2(t) \quad (t \leqslant T) \text{ or}$$
$$H_2\colon F_1(t) < F_2(t) \quad \text{or} \quad F_1(t) > F_2(t) \quad (t \leqslant T),$$

where F_1, F_2 are cumulative distributions (discrete or continuous) of the observations and T is their upper limit. The test is shown to be asymptotically normal and consistent against one-sided alternatives $F_1(t) < F_2(t)$ $(t \leqslant T)$ and against two-sided alternatives where either $F_1(t) < F_2(t)$ or $F_1(t) > F_2(t)$ $(t \leqslant T)$. The asymptotic efficiency of the test relative to the efficient parametric test when the distributions are exponential is at least 0·75 and increases with degree of censoring. When H_0 is true, the test is not seriously affected by real differences in the percentage censored in the two groups. Some comparisons are made for five cases of varying degrees of censoring and tying between probabilities from the exact test and those from the proposed test and these suggest the test is appropriate under certain conditions when the sample size is five in each group. A worked example is presented and some discussion is given to further problems.

2. Introduction

The statistical problem considered in this paper arises in clinical trials comparing two treatments, where the observation for each patient is often time to failure or censoring (sometimes referred to as loss). In fact, the results are relevant for distributions other than failure times and in fields of application outside medicine. However, the discussion is in terms of failure times since most applications are of this type and it is convenient to use medical terminology.

A common problem in a clinical trial is to compare two treatments for their ability to prolong life or maintain a patient in a well state. Patients enter study serially in time and are randomly allocated to one of two treatments. At a time T after the start of the study, an observation is recorded of time to failure (death or relapse) or censoring from observation (patient still alive or in remission at T). In general, $n_i - r_i$ individuals have failed and r_i are censored at time T $(i = 1, 2)$, but because patients have entered at different times, the times to censoring will differ among patients.

A special case has been considered by Halperin (1960) in an industrial life-testing context.

* This investigation was supported by a Public Health Service Research Grant from the National Cancer Institute.

Received by ER-RPF
SEP 13 1996
Dic

Los Alamos Environmental Restoration Records Processing Facility
11508

In his application, components could be started on test at the same time so that at the end of the experiment, times to censoring were the same for items not having failed. However, times to censoring could differ in industrial life-testing experiments where items are started at different times or where a policy of replacing failed items is followed.

If it is known that time to failure is exponentially distributed in both treatment groups, an F test can be used to test for treatment differences (see § 8). Since the exponential assumption is often not warranted and no other approach seems generally applicable, we consider a distribution-free two-sample test. The W test proposed is an extension of the Wilcoxon test to samples with arbitrary censoring on the right. The test is conditional on the given pattern of failures and censored observations.

Halperin (1960) and Rao, Savage & Sobel (1960) have considered two-sample tests with censoring, though all assume that times to censoring are the same in both samples. Recently, Alling (1963) has proposed a modified Wilcoxon test to be calculated sequentially so that an early decision may possibly be reached. His test is valid when censored observations are present, the test being based on least upper and greatest lower bounds for subsequent values of the ordinary Wilcoxon test statistic. The greatest saving in time of observation is when the sample sizes are small.

3. THE W STATISTIC AND RELATION TO OTHER STATISTICS

We assume that n_1, n_2 individuals are allocated randomly to treatments A, B, respectively, and we observe:

$$\left.\begin{array}{rl} x'_1, \ldots, x'_{r_1}, & r_1 \text{ censored} \\ x_{r_1+1}, \ldots, x_{n_1}, & n_1 - r_1 \text{ failures} \end{array}\right\} \text{ treatment } A,$$

$$\left.\begin{array}{rl} y'_1, \ldots, y'_{r_2}, & r_2 \text{ censored} \\ y_{r_2+1}, \ldots, y_{n_2}, & n_2 - r_2 \text{ failures} \end{array}\right\} \text{ treatment } B,$$

where x_i, y_j are times to failure and x'_i, y'_j are times to censoring (all measured from time of entry into study). It is emphasized again that the observations need not be failure times.

Such a pattern of observations could arise in a number of ways: in a clinical trial conducted for fixed time T where patients enter study serially in the interval 0 to T; in an industrial experiment where all components are started at time zero and an analysis is being done at time T later; in the same type of experiment, except that items that fail are replaced randomly; in a medical or industrial experiment where studies are being conducted at different centres, each study lasting a different length of time and an analysis is done by pooling results from all centres. Here T is the upper limit of time of observation among centres. A further possibility is a study of tolerances to different drugs when for some reason large tolerances cannot be measured accurately.

The test proposed is appropriate for these and possibly other cases with general types of censoring. The essential requirement is that the average exposure to the risk of failure be the same in the two groups. In other applications, the arbitrary censoring should be of the same type in both groups. In the sequel, the test is discussed in terms of the clinical trial, though it is clear that the other applications will also be relevant.

The times to failure are from cumulative distribution functions (C.D.F.'s) $F_1(x), F_2(y)$, which may be discrete or continuous. When considering the sample outcomes, we allow the possibility of ties among failure and loss times.

The null hypothesis is

$$H_0\colon F_1(t) = F_2(t) \quad (t \leqslant T) \quad \text{(treatments } A \text{ and } B \text{ equally effective).}$$

The alternative hypotheses (H_a) are either

$$H_1\colon F_1(t) < F_2(t) \quad (t \leqslant T) \quad \text{(treatment } A \text{ more effective than } B\text{)},$$

or the two-sided version

$$H_2\colon F_1(t) < F_2(t),$$

or

$$F_1(t) > F_2(t) \quad (t \leqslant T) \quad \text{(treatment } A \text{ or } B \text{ more effective)}.$$

Roughly, we are interested in one or two tail tests of the difference between the C.D.F.'s for treatments A and B.

We define

$$U_{ij} = \left.\begin{array}{lll} -1 & x_i < y_j & \text{or} \quad x_i \leqslant y_j', \\ 0 & x_i = y_j & \text{or} \quad (x_i', y_j') \quad \text{or} \quad x_i' < y_j \quad \text{or} \quad y_j' < x_i, \\ +1 & x_i > y_j & \text{or} \quad x_i' \geqslant y_j \end{array}\right\} \tag{3·1}$$

and calculate the statistic $W = \sum_{i,j} U_{ij}$ where the sum is over all $n_1 n_2$ comparisons. Hence, there will be a contribution to W for all comparisons of the two samples where both patients have failed (except for ties) and in all comparisons where a patient censored from observation has survived longer than one who has failed.

The W statistic is related to the Wilcoxon (1945) statistic T', the Mann–Whitney (1947) statistic U' and Kendall's (1955) statistic S when there are no censored observations or ties.

It is easy to show that

$$W = n_2(n_1+n_2+1) - 2T',$$

where T' is the sum of the ranks of the second sample in the ordered combined sample. Also,

$$W = 2U' - n_1 n_2,$$

where U' counts the number of times an observation in the second sample precedes one in the first in the combined ranking of the two samples. Further, $W = S$, a statistic defined by Kendall for use in rank correlation. The last is also true when ties are present.

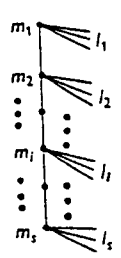
When all censored observations have the value T, Halperin's (1960) statistic U_c is defined by

$$W = 2U_c + r_1 r_2 - n_1 n_2,$$

where U_c is related to the Mann–Whitney statistic by $U_c = U'(n_1-r_1, n_2-r_2) + r_1(n_2-r_2)$. Here $U'(n_1-r_1, n_2-r_2)$ is the Mann–Whitney statistic based on the $n_1+n_2-r_1-r_2$ failures.

4. The conditional mean and variance of W

We have n_1, n_2 observations which can be arranged in the following general pattern:



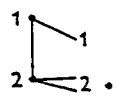
(4·1)

where

m_i = number of uncensored observations at rank i in rank ordering of uncensored observations with distinct values;

l_i = number of right-censored observations with values greater than observations at rank i but less than observations at rank $(i+1)$.

The points on the vertical line correspond to a rank ordering of the distinct values of the failure observations and these occur at s distinct failure points. Any set of failed and censored observations can be represented according to this diagram. If there are censored observations prior to the first failure, these could be included by counting them as l_1 with $m_1 = 0$. Ordinarily, such observations would be excluded since they provide no information on the differences between A and B. The calculation of mean and variance is not affected, since the calculation is conditional on the given pattern of observations. As an example, if a $+$ represents a censored observation and we have the following sample of survival times (weeks): 8, 8+, 10, 10, 11+, 14+, the pattern is



Suppose H_0 is true and the average exposure to the risk of failing is the same in the two groups. If the n_1+n_2 individuals in the pattern are labelled differently, there are $\binom{n_1+n_2}{n_1}$ possible allocations of the individuals to two samples with n_1, n_2 observations. We consider the conditional mean and variance of W under H_0. These are denoted by $E(W|P, H_0)$ and $\operatorname{var}(W|P, H_0)$, where P is the pattern of observations. The expectations are over the $(n_1+n_2)!/\{n_1!\,n_2!\}$ equally likely samples leading to the same observed pattern P.

It is easy to see

$$E(W|P, H_0) = 0, \tag{4.2}$$

by symmetry.

The derivation of the variance is given in Appendix A. The formula is

$$\operatorname{var}(W|P, H_0) = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)} \Big\{ \sum_{i=1}^{s} m_i M_{i-1}(M_{i-1}+1) + \sum_{i=1}^{s} l_i M_i (M_i+1) + \sum_{i=1}^{s} m_i (n_1+n_2-M_i-L_{i-1})(n_1+n_2-3M_{i-1}-m_i-L_{i-1}-1) \Big\}, \tag{4·3}$$

where

$$M_j = \sum_{i=1}^{j} m_i, \quad M_0 = 0,$$

$$L_j = \sum_{i=1}^{j} l_i, \quad L_0 = 0.$$

When there are no ties or losses, i.e. $m_1 = \ldots = m_s = 1$, $l_1 = \ldots = l_s = 0$, and $s = n_1+n_2$, the formula becomes

$$\operatorname{var}(W|P, H_0) = \tfrac{1}{3} n_1 n_2 (n_1+n_2+1)$$

which is the form expected from the variance of the Mann–Whitney (1947) statistic. Here, P is simply the ranking of the n_1+n_2 observations.

If there are no ties and all censored observations occur after the $(n_1+n_2-r_1-r_2)$th failure, i.e. $m_1 = \ldots = m_s = 1$, $l_1 = \ldots = l_{s-1} = 0$, $l_s = r_1+r_2$, and $s = n_1+n_2-r_1-r_2$ we have

$$\operatorname{var}(W|P, H_0) = \frac{n_1 n_2 (n_1+n_2-r_1-r_2)}{(n_1+n_2)(n_1+n_2-1)} \{(n_1+n_2)(r_1+r_2) + \tfrac{1}{3}[(n_1+n_2-r_1-r_2)^2-1]\},$$

which is that expected from the relation between W and the U_c of Halperin (1960).

Hemelrijk (1952) has given a formula for the variance of the Mann–Whitney statistic, U', allowing for ties. His formula gives the same result as (4·3) when there are tied and failure observations only.

When $m_1 = m, l_1 = l, m_i = l_i = 0\,(i \neq 1)$ (or equivalently $m_1 = m_1, m_2 = m_2, l_i = 0$ for all i, $m_i = 0\,(i > 2)$), the observations form a 2×2 contingency table with margins fixed and constitute an outcome in hypergeometric sampling. If two treatments are being compared in n_1, n_2 patients, the m individuals may be considered as the 'responders' having tied values on a response scale and the l individuals as 'non-responders', i.e. as being censored and requiring a greater stimulus to respond. The W statistic reduces to the difference in the products of the diagonals in the 2×2 table and

$$\operatorname{var}(W|P, H_0) = \frac{lmn_1 n_2}{(n_1+n_2-1)},$$

which is exactly the same as that obtained by assuming the W statistic to be an outcome in hypergeometric sampling.

5. The calculation of W and $\operatorname{var}(W|P, H_0)$ in large samples

This section can be conveniently skipped by those not concerned with the calculation of W in reasonably large samples (say n_1, n_2 both 25 or more); suffice to say that W and $\operatorname{var}(W|P, H_0)$ can be calculated quite easily by grouping the failure and censored observations. The W statistic and its variance are simple to calculate when n_1, n_2 are small. However if n_1, n_2 are large, then both the mean and variance calculation are lengthy.

Of course, it would not be difficult to program both calculations for an electronic computer. Alternatively, the failure and censored observations could be grouped in s intervals in a way similar to that of the life table:

	Treatment A		
Interval	No. of failures	Cum. no. of failures	No. of censored
1	f_{1A}	F_{1A}	c_{1A}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	f_{iA}	F_{iA}	c_{iA}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
s	f_{sA}	F_{sA}	c_{sA}

where

$$f_{iA} = \text{number of failures in interval } i,$$

$$c_{iA} = \text{number of censored observations in interval } i,$$

$$F_{iA} = \sum_{j=1}^{i} f_{jA}$$

and there is another table with entries f_{iB}, c_{iB} and F_{iB} defined in the same way for treatment B.

The intervals should be chosen the same as for ordinary frequency distributions and need not be of equal length. The failures in the ith interval are considered 'tied' at rank i in the rank ordering of intervals. The censored observations are also considered as 'tied' in the ith interval and are counted as occurring after interval $i-1$ but before i. Thus, information is lost concerning the ordering of failed and censored observations within each interval. The formula for W then becomes

$$W = \sum_{i=1}^{s} \{[f_{iA}+c_{iA}]F_{i-1,B} - [f_{iB}+c_{iB}]F_{i-1,A}\}, \tag{5·1}$$

where it is assumed that the same intervals are used in both samples. This statistic is quite simple to evaluate if each term is calculated successively by interval in the appropriate table.

The conditional variance of W is found by using the general formula (4·3) with

$$m_i = f_{iA} + f_{iB}, \quad l_i = c_{i+1,A} + c_{i+1,B}. \tag{5·2}$$

Both $E(W|P, H_a)$ (in absolute value) and $\mathrm{var}(W|P, H_0)$ will tend to be smaller on the average for the grouped case than for the ungrouped case. This results from the loss of a proportion of the $n_1 n_2$ comparisons because of grouping. If this proportion is not large, the test of W should not be seriously affected. In any doubtful case, the test on the ungrouped data could be carried out.

6. Asymptotic normality of W

In Appendix B, it is shown that W is asymptotically normal with mean and variance under the null hypothesis given by (4·2) and (4·3). The result follows because $(n_1 n_2)^{-1} W$ has the form of a two-sample U statistic, defined by Lehmann (1951), and a convergence theorem of Cramér (1946) may be applied to prove asymptotic normality. It is assumed that unconditionally the pattern of observations has arisen in a random way from a probability distribution of times to entry into study (in a special case, all patients enter at time zero) and two probability distributions of times to failure.

Consequently, to test H_0 against either H_1 or H_2, a value of

$$Z = \frac{W}{\sqrt{\{\mathrm{var}(W|P, H_0)\}}}$$

is taken as asymptotically normal with zero mean and unit variance.

The normal approximation is somewhat better if a continuity correction is made, especially if the sample sizes are not large. In an application where there are no or relatively few tied and censored observations, a continuity correction of ± 1 should be made. Here the possible W scores will usually be two units apart. Otherwise, the continuity correction should be $\pm \frac{1}{2}$.

The adequacy of the normal approximation is investigated in § 10. The results indicate that the W test can be applied when sample sizes are as small as $n_1 = n_2 = 5$, as long as not more than six of the ten observations are involved in ties or censoring and there are at least five distinct failure points. In the special case $m_1 = m$, $l_1 = l$, $m_i = l_i = 0$ $(i \neq 1)$ where the observations form a 2×2 contingency table, the W test is equivalent to the test based on the normal approximation to the hypergeometric distribution. Pearson (1947) has shown that even for moderate sample sizes the normal approximation gives probabilities in close agreement with those from the hypergeometric distribution.

7. The consistency of the W test

We now consider the behaviour of the W test when the null hypothesis is not true. For this, we need $E(W|P, H_a)$ and a bound for $\mathrm{var}(W|P, H_a)$. The alternative hypothesis here is fixed, that is it does not depend on the sample size in each group. Just as in Appendix B, we assume that considered unconditionally $X_1, \ldots, X_{n_1}$ are independent random vector variables taking values $(x, 0)$ or $(x', 1)$ if the sample outcome is a failure, censored observ

tion, respectively. There is a similar assumption for $Y_1, \ldots, Y_{n_2}$. For the sake of simplicity we also assume $n_1 = n_2 = n$.

We have
$$E(W|P, H_a) = n^2\{\Pr(X_i > Y_j|P, H_a) - \Pr(X_i < Y_j|P, H_a)\}$$
and this is non-zero when
$$\Pr(X_i > Y_j) - \Pr(X_i < Y_j) \neq 0. \tag{7·1}$$

For the variance, we write
$$\begin{aligned} n^{-4}\operatorname{var}(W|P, H_a) &= n^{-4}E(W^2|P, H_a) - n^{-4}[E(W|P, H_a)]^2 \\ &= n^{-4}E\{\Sigma U_{ij}^2 + \sum_{i \neq i'} U_{ij}U_{i'j} + \sum_{j \neq j'} U_{ij}U_{ij'} + \sum_{\substack{i \neq i' \\ j \neq j'}} U_{ij}U_{i'j'}|P, H_a\} \\ &\quad - n^{-4}[E\{\Sigma U_{ij}|P, H_a\}]^2. \end{aligned}$$

The four terms comprising $E(W^2|P, H_a)$ have n^2, $n^2(n-1)$, $n^2(n-1)$ and $n^2(n-1)^2$ individual terms in the summations, each with expectation at most one. Therefore, the first three terms in the above expression are at most $O(1/n)$. But
$$\begin{aligned} E\{\sum_{\substack{i \neq i' \\ j \neq j'}} U_{ij}U_{i'j'}|P, H_a\} &= \sum_{\substack{i \neq i' \\ j \neq j'}} E(U_{ij}|P, H_a)\,E(U_{i'j'}|P, H_a) \\ &= n^2(n-1)^2\,[E(U_{ij}|P, H_a)]^2 \end{aligned}$$
and
$$\begin{aligned} [E\{\Sigma U_{ij}|P, H_a\}]^2 &= [\Sigma E(U_{ij}|P, H_a)]^2 \\ &= n^4[E(U_{ij}|P, H_a)]^2 \end{aligned}$$
so that $n^{-4}\operatorname{var}(W|P, H_a)$ is at most $O(1/n)$.

Therefore
$$\lim_{n \to \infty} n^{-4}\operatorname{var}(W|P, H_a) = 0.$$

Hence
$$\frac{n^{-2}E(W|P, H_a) - n^{-2}E(W|P, H_0)}{\sqrt{\{n^{-4}\operatorname{var}(W|P, H_a)\}}}$$
becomes indefinitely large as $n \to \infty$ and the probability that the W test rejects the null hypothesis is 1.

Thus, the W test is consistent for alternatives where (7·1) is true. In particular, the one-sided W test is consistent against alternatives $F_1(t) < F_2(t)$ $(t \leqslant T)$ and the two-sided test against alternatives where either $F_1(t) > F_2(t)$ or $F_1(t) < F_2(t)$ $(t \leqslant T)$.

8. The asymptotic relative efficiency of W to F assuming exponential failure distributions

Suppose the probability density function of time to failure for a patient receiving treatment A is
$$f_1(x) = \phi \exp(-\phi x)$$
and that for a patient receiving treatment B is
$$f_2(y) = \theta\phi \exp(-\theta\phi y).$$

We wish to test the hypothesis
$$H\colon F_1(t) = F_2(\theta t) \quad (t \leqslant T, 0 < \theta \leqslant 1),$$
where under H_0: $\theta = 1$ and under H_1: $0 < \theta < 1$. Such a test would be appropriate if we were interested in whether failure times on treatment B were a constant proportion (θ) of those on treatment A. For example, if $\theta = 0{\cdot}75$, the failure times for the patients on

treatment A would be 25 % longer than those on treatment B. A test of the above hypothesis is equivalent to one for differences in location, $F_1(t) \doteq F_2(t+\theta)$, when logs of failure times are analysed.

An efficient parametric test for the hypothesis is to take $\bar{t}_1/\bar{t}_2$ as having an F distribution with $(2(n_1-r_1), 2(n_2-r_2))$ degrees of freedom, where

$$\bar{t}_1 = \left(\sum_{i=1}^{r_1} x_i' + \sum_{i=r_1+1}^{n_1} x_i\right)\Big/(n_1-r_1), \quad \bar{t}_2 = \left(\sum_{i=1}^{r_2} y_i' + \sum_{i=r_2+1}^{n_2} y_i\right)\Big/(n_2-r_2).$$

The F distribution is exact when the time of observation in each group is a random variable and n_1-r_1, n_2-r_2 are fixed, and a good approximation (Cox, 1953) when the time of observation is fixed and the number of patients failing prior to that is random.

We wish to calculate the asymptotic efficiency of the generalized Wilcoxon test relative to the F test in two situations:

(a) all individuals enter study at time zero, observation stops at T (the case where all individuals fail is covered by letting $T \to \infty$),

(b) individuals enter study at a constant rate, λ, in the interval 0 to T and fail according to $f_1(x)$ or $f_2(y)$.

For both cases, it is assumed that the number of patients in each group is n. Case (b) is a model of a clinical trial, also suggested by Armitage (1959), where it is reasonable to assume there is a fixed probability, $\lambda(\Delta t)$, of a patient entering a study in any small interval of time (Δt). Unconditionally both the number of patients entering study and the total time of exposure to the risk of failing are random variables. Conditional on $2n$ patients being entered in 0 to T, the times of entry will be distributed independently and uniformly over the interval $(0, T)$.

For case (a), the chance of an individual being censored at time T is $e^{-T\phi}$, $e^{-T\theta\phi}$ for those receiving treatments A, B, respectively. For case (b), the same chances are $(1-e^{-T\phi})/(T\phi)$ and $(1-e^{-T\theta\phi})/(T\theta\phi)$. Further details are given in Appendix C.

In the calculation, it is convenient to transform the F statistic to $z = \frac{1}{2}\log F$ so that z is asymptotically normal with

$$\operatorname{var}(z) \cong \frac{1}{2}\left(\frac{1}{2(n_1-r_1)} + \frac{1}{2(n_2-r_2)}\right),$$

where $2(n_1-r_1)$, $2(n_2-r_2)$ are the number of degrees of freedom in F. Also, we arrange that the variance of each test statistic is of order n^{-1} by considering $n^{-2}W$ rather than W.

To obtain an asymptotic measure of test efficiency, we consider a sequence of alternative hypotheses in which θ approaches the value tested, $\theta = 1$, as n increases. In this case, the asymptotic efficiency of W relative to F is

$$\text{A.R.E.} = \lim_{n\to\infty} \frac{\left\{\left.\dfrac{\partial E(n^{-2}W)}{\partial\theta}\right|_{\theta=1}\right\}^2}{\{n \operatorname{var}(n^{-2}W|H_0)\}} \times \frac{\{n \operatorname{var}(z|H_0)\}}{\left\{\left.\dfrac{\partial E(z)}{\partial\theta}\right|_{\theta=1}\right\}^2} \tag{8·1}$$

and the calculation of the terms required is given in Appendix C. A good exposition of the concept of asymptotic relative efficiency (A.R.E.) is given by Kendall & Stuart (1961, pp. 265–76).

Values of A.R.E. for case (a) and (b) are given in Table 1 for various values of $T\phi$.

Now

$$T\phi = \frac{\text{total study time}}{\text{average failure time on treatment } A},$$

that if an estimate of $T\phi$ is available, some idea of the A.R.E. can be obtained. Note that the minimum value of A.R.E. is 0·75 for both cases (a) and (b) and that A.R.E. increases as $T\phi$ decreases. As $T\phi \to 0$, A.R.E. approaches one for case (a) and 8/9 for case (b). Clinical trials are often conducted with $T\phi$ about 2 or 3 and here A.R.E. is close to 0·80.

Table 1. *Asymptotic efficiency of W relative to F assuming exponential failure in two groups*

Treatment A: $\phi \exp(-\phi x)$; treatment B: $\theta\phi \exp(-\theta\phi y)$

Case (a). All individuals enter study at time zero, observation stops at T.

	$T\phi \to \infty$	$T\phi = 3$	$T\phi = 2$	$T\phi = 1$	$T\phi \to 0$
A.R.E.	0·750	0·785	0·838	0·934	1

Case (b). Individuals enter study according to uniform distribution over $(0, T)$ and study stops at T.

	$T\phi \to \infty$	$T\phi = 3$	$T\phi = 2$	$T\phi = 1$	$T\phi \to 0$
A.R.E.	0·750	0·781	0·802	0·836	8/9

A rationale for these results is as follows: consider the patterns of observations for a case (a) situation with a high degree of censoring. The patterns might appear as

The ratio of the means in the F test will differ from one when there is a difference in the *number* of failures and *times* to failure between A and B. The value of W depends mainly on the difference in the *number* of failures between groups. The A.R.E. result means that as $T\phi$ becomes small and there is more censoring, the *times* to failure are not much more important than the *number* of failures. The same type of result was found by Armitage (1959) for paired data, when he compared the A.R.E. of the sign method to parametric maximum likelihood for exponential distributions.

The increase in A.R.E. is slower for case (b) as $T\phi \to 0$. In this situation, it would be expected that those individuals censored would be among the later entrants to study and so would tend to occur near the beginning of the pattern. Knowledge of the times to failure would then be relatively more important.

These results suggest that the W test would be reasonable to apply when comparing failure time distributions, especially when some censoring is expected. When the distributions are not exponential, a two-parameter distribution such as the Weibull might be needed. If X has a Weibull distribution, it is well known that $X^{1/\alpha}$ has an exponential distribution. But any such power transformation would not affect W since W is rank invariant. Hence, the calculations of A.R.E. given would be exactly the same if the distributions were assumed to be Weibull with *known* index. It is reasonable to suppose that the A.R.E. values would be at least as great if the distributions were Weibull with *unknown* index, so that the stated values are lower bounds for A.R.E.

9. LOSS RATE DIFFERENT IN THE TWO GROUPS

Suppose that in an application there is in fact no difference in the C.D.F.'s of times to failure, but that for some reason there is a difference in the percentage censored in the two groups. In an extreme case, all individuals are observed to failure in one group and study stops at time T_1 in the other group. This could happen in a clinical trial if the drug given to patients in one group had deteriorated by time T_1 or it was not possible to administer treatment after time T_1. In such cases, it would only be appropriate to consider failures and censored observations up to time T_1 in the affected group. We assume that the sample size is n in each group and all individuals have entered study at time zero, so T_1 is the length of study for all individuals in the affected group.

Table 2. *Ratio of* $\sqrt{\{\mathrm{var}\,(W|H_0)\}}$ *to* $\sqrt{\{\mathrm{var}_R\,(W|H_0)\}}$ *for various* $1-r/n$

	$1-r/n$						
Ratio	1	0·95	0·9	0·8	0·7	0·6	0·1
$\dfrac{\sqrt{\{\mathrm{var}\,(W\|H_0)\}}}{\sqrt{\{\mathrm{var}_R\,(W\|H_0)\}}}$	1	1·035	1·065	1·115	1·152	1·180	1·357

What happens to the mean and variance of W if an analysis is performed without considering all censored observations restricted to one group? We are concerned with the average effect of censoring one group at T_1 and so we calculate $E(W|H_0)$ and $\mathrm{var}\,(W|H_0)$, that is we average over the possible patterns that could occur. When there are n patients per group and H_0 is true, the number failing in each prior to T_1 is binomially distributed with expected value $n-r$ and variance

$$n\left(\frac{r}{n}\right)\left(1-\frac{r}{n}\right).$$

The means and variances for the two cases are as follows (taking terms to $O(1/n)$ for the variances):

Losses restricted to one group

$$E_R(W|H_0) = 0,$$

$$n^{-4}\,\mathrm{var}_R\,(W|H_0) \cong \frac{2}{3}\left(1-\frac{r}{n}\right)^3\frac{1}{n}+\frac{2}{n}\left(\frac{r}{n}\right)\left(1-\frac{r}{n}\right).$$

Losses not restricted

$$E(W|H_0) = 0,$$

$$n^{-4}\,\mathrm{var}\,(W|H_0) \cong \frac{1}{12n}+\frac{11}{4}\left(\frac{r}{n}\right)\left(1-\frac{r}{n}\right)+\frac{7}{12n}\left(1-\frac{r}{n}\right)^3.$$

Thus, if the mean and variance of W are calculated in the usual way, there is no bias in the estimate of the mean but the estimate of variance will be an over-estimate. To examine the extent of the over-estimate, Table 2 gives the ratio of the two standard errors for various values of $1-r/n$.

The ratio of the standard errors is less than 1·2 even when 40 % of the observations are censored at T_1 in the affected group. There will be some loss in sensitivity in detecting departures from the null hypothesis when the ordinary W test is applied; however, this is unlikely to be serious when the proportion of censored observations is under about 20 %.

10. Some exact calculations

Table 3 presents a comparison of tail probabilities using the generalized Wilcoxon test (corrected for continuity) and exact calculation for five cases of varying degrees of censoring and tying. The sample size in each group is $n_1 = n_2 = 5$ and the pattern of observations for each case is given. The cases were selected arbitrarily to represent various degrees of censoring and tying. The total number of observations involved in ties or censoring ranges from four for case IV to nine for case II.

For each case, the first column gives the cumulative frequency of a given score or larger from the exact distribution. It was necessary to evaluate scores from only $\frac{1}{2}\binom{10}{5} = 126$ possible samples, since the distribution of W is symmetric. The second column gives the exact probability of a given score or larger and the last column gives the estimated probability using the W test, corrected for continuity by subtracting $\frac{1}{2}$ from each score.

The probabilities from the W test are remarkably close to the exact probabilities considering the small sample sizes and heavy tying and censoring. This is especially so in the tail of the distribution where most interest lies. The approximation is poorest when a large number of samples have the same score, but this happens more often near the centre of the distribution. If it is decided arbitrarily that the normal approximation is adequate when the absolute difference between approximate and exact is 0·01 or less up to a cumulative tail probability of 0·10, then only cases II and III fail to satisfy this criterion. In both of these cases, the number of different failure points is only four and the total number of observations involved in ties or censoring is nine and six, respectively.

It is difficult to make a general statement on the sizes of sample necessary before the asymptotic theory holds because of the varying degrees of censoring and tying that are possible. Halperin (1960), for his case, has stated that when $n_1 = n_2 = 8$ the asymptotic normal theory is adequate for all practical purposes up to about 75 % censoring (no tying) at both the 5 and 1 % significance levels. Lehman (1961) considered the exact and approximate distributions of the Wilcoxon statistic when $n_1 = n_2 = 5$ for five cases of varying degrees of tying. Using the arbitrary criterion above at significance levels of 0·01, 0·05, and 0·10, the normal approximation was adequate when the number of observations involved in ties was six or less. Taking the results here with the others, the normal approximation with continuity correction seems adequate when $n_1 = n_2 = 5$, as long as the total number of observations tied or censored is six or less and there are at least five distinct failure points.

Of course, if the application of the W test is doubtful in a particular case, the following rule seems reasonable: calculate W (corrected for continuity) and if the result is borderline (say 0·03 to 0·10), calculate the exact test. Otherwise accept the verdict of the W test.

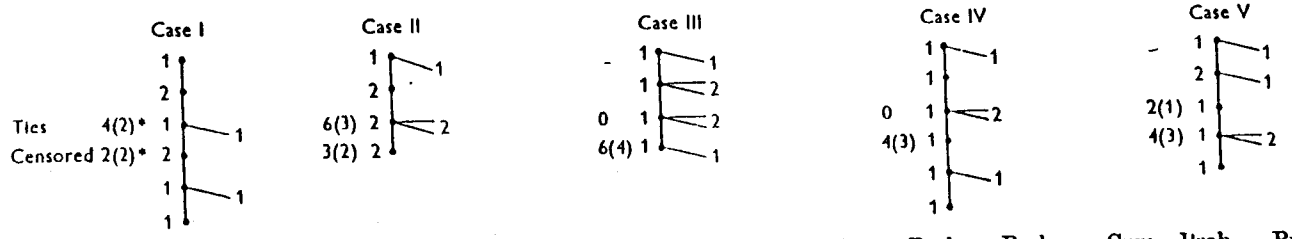
11. A worked example

In this section, we apply the W test to an example from a clinical trial. In the trial, reported by Freireich *et al.* (1963), 6-mercaptopurine (6-MP) was compared to a placebo in the maintenance of remissions in acute leukemia. The trial was actually conducted sequentially, but will be here analysed as a fixed sample size trial. One year after the start of the study, the following lengths of remission were recorded:

Length of remission (weeks)

6-MP (21)	6, 6, 6, 7, 10, 13, 16, 22, 23 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+
Placebo (21)	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Table 3. *Observation patterns and probabilities of scores by exact calculation and the generalized Wilcoxon test (corrected for continuity) for 5 cases of censoring and tying in samples of size $n_1 = 5$, $n_2 = 5$*



Score	Case I Cum. freq.	Case I Prob. (exact)	Case I Prob. (W_c)	Case II Cum. freq.	Case II Prob. (exact)	Case II Prob. (W_c)	Case III Cum. freq.	Case III Prob. (exact)	Case III Prob. (W_c)	Case IV Cum. freq.	Case IV Prob. (exact)	Case IV Prob. (W_c)	Case V Cum. freq.	Case V Prob. (exact)	Case V Prob. (W_c)
23	2	0·0079	0·0064	.	.	.	.	.	.	.	.	.	.	.	.
22	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
21	.	.	.	1	0·0040	0·0055	.	.	.	.	.	.	.	.	.
20	4	·0159	·0154	.	.	.	.	.	.	.	.	.	2	0·0079	0·0043
19	5	·0198	·0202	3	·0119	·0110	.	.	.	1	0·0040	0·0104	.	.	.
18	.	.	.	.	.	.	.	.	.	2	·0079	·0146	4	·0159	·0170
17	7	·0278	·0336	.	.	.	.	.	.	5	·0198	·0202	.	.	.
16	13	·0516	·0427	.	.	.	.	.	.	7	·0278	·0274	9	·0357	·0314
15	.	.	.	13	·0516	·0359	.	.	.	11	·0436	·0375	.	.	.
14	.	.	.	.	.	.	3	0·0119	0·0239	14	·0556	·0495	15	·0595	·0537
13	27	·1071	·0838	17	·0675	·0606	8	·0317	·0336	17	·0675	·0643	.	.	.
12	.	.	.	.	.	.	15	·0595	·0465	23	·0913	·0823	22	·0873	·0885
11	.	.	.	26	·1032	·0968	19	·0754	·0619	26	·1032	·1038	.	.	.
10	45	·1786	·1469	.	.	.	24	·0952	·0824	34	·1349	·1314	40	·1587	·1379
9	49	·1944	·1736	48	·1905	·1469	31	·1230	·1076	39	·1548	·1611	.	.	.
8	.	.	.	.	.	.	41	·1627	·1357	52	·2063	·1949	54	·2143	·2004
7	63	·2500	·2358	52	·2063	·2090	49	·1944	·1712	61	·2420	·2327	56	·2222	·2389
6	77	·3055	·2709	.	.	.	58	·2301	·2120	75	·2976	·2743	79	·3135	·2810
5	.	.	.	88	·3492	·2878	66	·2619	·2546	87	·3452	·3228	81	·3214	·3264
4	83	·3293	·3483	.	.	.	76	·3016	·3050	98	·3889	·3707	101	·4008	·3745
3	111	·4404	·3808	100	·3968	·3783	87	·3452	·3657	113	·4484	·4207	.	.	.
2	.	.	.	.	.	.	101	·4008	·4129	121	·4801	·4721	126	·5000	·4761
1	112	·4444	·4801	126	·5000	·4761	118	·4682	·4721	131	·5198	·5279	.	.	.
0	126	·5000	·5199	.	.	.	134	·5317	·5279	.	.	.	.	.	.
$\sqrt{\mathrm{var}(W\|P, H_0)}$	9·04			8·06			6·83			7·50			7·78		

* Number in ( ) is number of distinct points.

A + sign indicates a censored observation. Because the upper limit of observation time is about 35 weeks, $T = 35$ weeks. In this case, it is clear that 6-MP is the superior treatment, there being 12 censored observations at long remission times on 6-MP and none on placebo. The data will be analysed to illustrate the calculations.

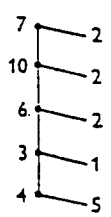
To calculate W directly, a 21×21 table is formed with the failures and censored observations in each group ordered separately along the margins. Entries of $+1$, -1 or 0 are made in accordance with the scoring scheme (3·1) for the 441 comparisons. The result is $W = 335 - 64 = 271$. Since $n_1 \times n_2$ is rather large, it is natural to consider the result obtained by grouping observations. Then W is obtained from the formula for grouped data given by (5·1) and can be calculated conveniently in the format:

Interval (weeks)	6-MP f_{iA}	6-MP F_{iA}	6-MP c_{iA}	6-MP a_i	Placebo f_{iB}	Placebo F_{iB}	Placebo c_{iB}	Placebo b_i
0– 4	0	0	0	—	7	7	0	—
5– 9	4	4	2	42	6	13	0	0
10–14	2	6	2	52	4	17	0	16
15–19	1	7	2	51	2	19	0	12
20–24	2	9	1	57	2	21	0	14
25–	0	9	5	105	0	21	0	0
	9		12	307	21		0	42

Thus
$$W = \sum_{i=1}^{6} \{a_i - b_i\} = 307 - 42 = 265,$$
where
$$a_i = [f_{iA} + c_{iA}]F_{i-1,B}, \quad b_i = [f_{iB} + c_{iB}]F_{i-1,A}.$$

The var $(W|P, H_0)$ for the grouped data is obtained from (4·3) with
$$m_i = f_{iA} + f_{iB}, \quad l_i = c_{i+1,A} + c_{i+1,B}.$$

The pattern is



and the format for calculating the variance is

i	m_i	M_i	d_i	$m_i \times d_{i-1}$	l_i	L_{i-1}	$l_i \times d_i$	e_i	f_i	$e_i \times f_i$	$m_i \times e_i \times f_i$
1	7	7	56	0	2	0	112	35	34	1190	8330
2	10	17	306	560	2	2	612	23	8	184	1840
3	6	23	552	1836	2	4	1104	15	−20	−300	−1800
4	3	26	702	1656	1	6	702	10	−37	−370	−1110
5	4	30	930	2808	5	7	4650	5	−48	−240	−960
	30			6860	12		7180				6300

where
$$d_i = M_i(M_i + 1), \quad d_0 = 0,$$
$$e_i = n_1 + n_2 - M_i - L_{i-1},$$
$$f_i = n_1 + n_2 - 3M_{i-1} - m_i - L_{i-1} - 1.$$

Then,
$$\operatorname{var}(W|P, H_0) = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)}\left\{\sum_{i=1}^{5} m_i d_{i-1} + \sum_{i=1}^{5} l_i d_i + \sum_{i=1}^{5} m_i e_i f_i\right\}$$
$$= \frac{(21)(21)}{(42)(41)}\{6860+7180+6300\} = 5065{\cdot}6,$$
$$\sqrt{\{\operatorname{var}(W|P, H_0)\}} = 71{\cdot}2.$$

The result obtained from the ungrouped data is $\sqrt{\{\operatorname{var}(W|P, H_0)\}} = 75{\cdot}1$.

Suppose we wish to test H_0: $F_1(t) = F_2(t)$ $(t \leqslant T)$ against the alternative H_2: $F_1(t) < F_2(t)$ or $F_1(t) > F_2(t)$ $(t \leqslant T)$. We are interested in whether 6-MP lengthens or shortens remissions relative to placebo. We calculate
$$Z = \frac{W}{\sqrt{\{\operatorname{var}(W|P, H_0)\}}} = \frac{265}{71{\cdot}2} = 3{\cdot}72$$
and the probability of such a value of Z or a larger one in absolute value is about 0·0002 from tables of the normal distribution. Consequently there is very strong evidence that patients receiving 6-MP have longer remissions than those receiving placebo.

If the test is done with the ungrouped data, we find $Z = 3{\cdot}61$ and $\Pr(Z) \approx 0{\cdot}0004$. The result is quite close to that for the grouped data considering the moderate sample sizes in each group.

12. DISCUSSION

Some further problems connected with the generalized Wilcoxon test are: the extension of the test to the case of double censoring (i.e. in the upper and lower tails of the variable),* the extension of the test to more than two samples,* the development of a sequential W test and the use of the W test to find confidence limits.

In principle, there is no difficulty in extending the W test to the case of double censoring. The pattern of observations given by (4·1) could be generalized by considering l_i individuals $(i = 1, \ldots, s)$ to be censored on the left at a point immediately prior to the failure of the m_i individuals at rank i in the ordering of distinct failures. The change in the scoring of W given by (3·1) would be simple using the ordering relationships in the generalized pattern, the assumption being made that individuals censored on the left or right cannot be ordered among themselves. The proofs of asymptotic normality and consistency of the test based on W follow directly from those given here.

The extension of the W test to the k-sample case could be made in a way analogous to that suggested by Terpstra (1952) and Jonckheere (1954) for the extension of the ordinary Wilcoxon test. The null hypothesis is that all samples come from the same population and this is to be tested against the ordered alternative hypothesis: $F_1(t) < F_2(t) < \ldots < F_k(t)$. Suppose the statistic W is calculated for all $\frac{1}{2}k(k-1)$ pairs of samples. If we write W_{pq} for the value obtained from the pth and qth samples $(p, q = 1, 2, \ldots, k;\ p \neq q)$, then we can consider
$$W_k = \sum_{p=1}^{k} \sum_{q=p+1}^{k} W_{pq}.$$
From the results of Terpstra and Jonckheere, the limiting distribution of W_k should be normal.

* I am indebted to Professor J. Hemelrijk and a referee for helpful comments concerning these extensions.

There are at least two ways to consider a sequential W test. First, suppose an experiment comparing failure times is set up and n_1, n_2 items are placed on test in each group. The problem is to devise a test to stop the experiment at the earliest *time* possible (no saving in number of observations). One solution for this problem has been proposed by Alling (1963) based on least upper and greatest lower bounds for subsequent values of the ordinary Wilcoxon test statistic. The W test could be applied sequentially in time but the conditions necessary for this require investigation. Alternatively, it would often be desirable to conduct a sequential experiment that may result in a saving of time and observations. For example, suppose a clinical trial is being conducted and the hypothesis being tested is of type $F_1(t) = F_2(\theta t)$ with different values of θ specified for alternative hypotheses. Individuals are entered sequentially in each group and some form of W test is carried out sequentially in time. Under what conditions could such a sequential experiment be carried out?

Approximate confidence limits for the scale parameter θ can be found using the W statistic when the model is $F_1(t) = F_2(\theta t)$. The idea is to obtain an estimate of the confidence limits for θ assuming an underlying exponential distribution and then use the W test to find the approximate level of confidence for the limits. Thus the confidence limits are distribution free; the exponential assumption is introduced merely to get convenient starting values. If the failure time distributions in the two groups are exponential, then $\bar{t}_1/\bar{t}_2$ as defined in § 8 is an estimate of θ and confidence limits can be derived from the F distribution. All observations in the second sample are multiplied by the upper and lower confidence limits for θ and two W tests are carried out using the new values for the second sample. Two normal deviates will be obtained, say Z_1, Z_2, and the approximate level of confidence that θ lies between these limits can be calculated from tables of the normal distribution. For example, using the data of § 11 with those receiving placebo as group 2, we find $\bar{t}_1/\bar{t}_2 = 39{\cdot}9/8{\cdot}7 = 4{\cdot}6$ and 95 % confidence limits for θ: $1{\cdot}9 < \theta < 10$. After two W tests, we estimate that the level of confidence for these limits is about 92 %. Generally, the distribution-free confidence intervals will be wider than the corresponding intervals when the exponential assumption is made.

I wish to thank Professor D. R. Cox for very helpful suggestions and encouragement throughout the course of this work. Also, I want to thank my wife, Brenda, for doing the calculations in § 10.

References

Alling D. (1963). Early decision in the Wilcoxon two-sample test. *J. Amer. Statist. Ass.* 58, 713–20.

Armitage, P. (1959). The comparison of survival curves. *J.R. Statist. Soc.* A, 122, 279–300.

Cox, D. R. (1953). Some simple approximate tests for Poisson variates. *Biometrika*, 40, 354–60.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Freireich, E. J. *et al.* (1963). The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia. *Blood*, 21, 699–716.

Halperin, M. (1960). Extension of the Wilcoxon–Mann–Whitney test to samples censored at the same fixed point. *J. Amer. Statist. Ass.* 55, 125–38.

Hemelrijk, J. (1952). Note on Wilcoxon's two-sample test when ties are present. *Ann. Math. Statist.* 23, 133–5.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* 19, 293–325.

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41, 133–45.

Kendall, M. G. (1955). *Rank Correlation Methods*, 2nd edn. London: Charles Griffin and Co.

KENDALL, M. G. & STUART, A. (1961). *The Advanced Theory of Statistics*, 2. London: Charles Griffin and Co.

LEHMAN, S. (1961). Exact and approximate distributions for the Wilcoxon statistic with ties. *J. Amer. Statist. Ass.* 56, 293–8.

LEHMANN, E. (1951). Consistency and unbiasedness of certain non-parametric tests. *Ann. Math. Statist.* 22, 165–79.

MANN, H. B. & WHITNEY, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.* 18, 50–60.

PEARSON, E. S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table. *Biometrika*, 34, 139–67.

RAO, V., SAVAGE, I. R. & SOBEL, M. (1960). Contributions to the theory of rank order statistics: the two-sample censored case. *Ann. Math. Statist.* 31, 415–26.

TERPSTRA, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Proc. Kon. Ned. Akad. Wetensch.* A, 55, 327–33.

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80–3.

APPENDIX A

The variance of W under H_0 and conditional on a given pattern (P) of failed and censored observations is

$$\operatorname{var}(W|P,H_0) = E\{\sum_{i,j} U_{ij} - E(\sum_{i,j} U_{ij})|P,H_0\}^2.$$

The expectation is over the $(n_1+n_2)!/(n_1!\,n_2!)$ equally likely samples from the same pattern (general form is given by (4·1)). This may be written

$$\operatorname{var}(W|P,H_0) = E\{\sum_{i,j} U_{ij}^2 + \sum_{i \neq i'} U_{ij}U_{i'j} + \sum_{j \neq j'} U_{ij}U_{ij'} + \sum_{\substack{i \neq i' \\ j \neq j'}} U_{ij}U_{i'j'}|P,H_0\}, \tag{A 1}$$

since $E\{\sum_{i,j} U_{ij}|P,H_0\} = 0$, by symmetry.

We now proceed to evaluate each term in (A 1). We have

$$E\{\sum_{i,j} U_{ij}^2|P,H_0\} = \frac{2\binom{n_1+n_2-2}{n_1-1}}{\binom{n_1+n_2}{n_1}}\left\{\sum_{i=1}^{s}\left[\binom{m_i}{1}\binom{M_{i-1}}{1}+\binom{l_i}{1}\binom{M_i}{1}\right]\right\}, \tag{A 2}$$

where the term outside the brackets is the proportion of times a particular pair (i,j) will occur in opposite samples. The first term in the brackets is the number of ways of pairing a failed observation at rank i with one of lower rank and the second term is the number of ways of pairing an observation censored just after rank i with one having failed earlier.

Also,

$$E\{\sum_{i \neq i'} U_{ij}U_{i'j}|P,H_0\} = \frac{2\binom{n_1+n_2-3}{n_1-2}}{\binom{n_1+n_2}{n_1}}\{K\}, \tag{A 3}$$

where
$$K = \sum_{i=1}^{s}\binom{m_i}{1}\left[\binom{M_{i-1}}{2}+\binom{n_1+n_2-M_i-L_{i-1}}{2}\right] + \sum_{i=1}^{s}\binom{l_i}{1}\binom{M_i}{2} - \sum_{i=1}^{s}\binom{m_i}{1}\binom{M_{i-1}}{1}\binom{n_1+n_2-M_i-L_{i-1}}{1}.$$

The outside term in (A 3) is the proportion of times a particular pair of observations (i,i') will occur in one sample and a particular observation (j) in the other sample. The first term in K gives the number of ways of finding a meaningful pair (i,i') below and above j when j is a failure observation. The second term gives the number of ways of finding a pair of failure observations (i,i') of lower rank than j when j is a censored observation. The last term is the number of ways of finding one observation above and one failure observation below j when j is a failure.

Now,

$$E\{\sum_{j \neq j'} U_{ij}U_{ij'}|P,H_0\} = \frac{2\binom{n_1+n_2-3}{n_2-2}}{\binom{n_1+n_2}{n_1}}\{K\}, \tag{A 4}$$

by symmetry. Finally,

$$E\{\sum_{\substack{i \neq i' \\ j \neq j'}} U_{ij}U_{i'j'}|P,H_0\} = 0$$

because U_{ij} is independent of U_{rs} and $E(U_{ij}|P,H_0) = E(U_{rs}|P,H_0) = 0$. Collecting terms in (A 1) using (A 2), (A 3) and (A 4) and simplifying, we have $\mathrm{var}(W|P,H_0)$ as given by (4·3).

Appendix B

To show that W is asymptotically normal, we note first that $(n_1 n_2)^{-1}W$ has the form of a two-sample U statistic. Lehmann (1951) proved that such statistics are asymptotically normal using a general theorem of Hoeffding (1948). We give a definition of a two-sample U statistic sufficient for our purposes:

Let $X_1, \ldots, X_{n_1}; Y_1, \ldots, Y_{n_2}$ be n_1+n_2 independent, random vectors $X_\alpha = (X_\alpha^{(1)}, X_\alpha^{(2)})$, $Y_\beta = (Y_\beta^{(1)}, Y_\beta^{(2)})$, with cumulative distribution functions (c.d.f.) $F_1(x), F_2(y)$ where $x_\alpha = (x_\alpha^{(1)}, x_\alpha^{(2)})$ and $y_\beta = (y_\beta^{(1)}, y_\beta^{(2)})$. For $n_1, n_2 \geqslant 1$ and a real valued function defined by $t(X_\alpha, Y_\beta)$, a statistic

$$U = \frac{1}{n_1 n_2} \sum_{\alpha,\beta} t(X_\alpha, Y_\beta) \quad \begin{pmatrix} \alpha = 1, \ldots, n_1 \\ \beta = 1, \ldots, n_2 \end{pmatrix} \tag{B 1}$$

is a two-sample U statistic. Lehmann (1951) showed U to be asymptotically normal when $n_1 \to \infty$ such that $\lim n_1/n_2$ exists and under conditions that $E\{t(X_\alpha, Y_\beta)\} = \eta$ and $E\{t(X_\alpha, Y_\beta)\}^2 = M < \infty$.

The difficulty with applying these results directly to the W statistic is that the distribution of W has been considered *conditionally* for a given pattern of failed and censored observations and so we do not have n_1+n_2 independent random variables. However, we can show that, considered unconditionally, $(n_1 n_2)^{-1}W$ is a two-sample U statistic and then apply a convergence theorem to prove asymptotic normality.

Suppose there is a probability distribution of times to entry of the n_1+n_2 patients entering study in the interval 0 to T. This distribution may be of a very general type: a discrete lump of probability with all patients entering at time 0, a uniform distribution, or various distributions with a bunching of patients near time 0. The only assumption is that the distribution of patient entries is such that the number of failures at time T becomes large as n_1, n_2 become large.

Now define

$$x_\alpha = (x_\alpha^{(1)}, x_\alpha^{(2)}) \quad (\alpha = 1, \ldots, n_1),$$

where $x_\alpha^{(1)} = x_i, x_i'$ (time to failure, censoring) is from $F_1(x_\alpha^{(1)})$ and $x_\alpha^{(2)}$ is an indicator taking a value 0, 1 as $x_\alpha^{(1)}$ is a time to failure, censoring. A similar set-up is defined for y_β. Then, $X_1, \ldots, X_{n_1}; Y_1, \ldots, Y_{n_2}$ are n_1+n_2 independent random vectors.

If we now define

$$t(X_\alpha, Y_\beta) = \begin{cases} -1 & \text{if } x_\alpha^{(1)} < y_\beta^{(1)} \text{ and } (x_\alpha^{(2)}, y_\beta^{(2)}) \text{ is } (0,0), \\ & \text{or } x_\alpha^{(1)} \leqslant y_\beta^{(1)} \text{ and } (x_\alpha^{(2)}, y_\beta^{(2)}) \text{ is } (0,1), \\ 0 & \text{otherwise}, \\ +1 & \text{if } x_\alpha^{(1)} > y_\beta^{(1)} \text{ and } (x_\alpha^{(2)}, y_\beta^{(2)}) \text{ is } (0,0), \\ & \text{or } x_\alpha^{(1)} \geqslant y_\beta^{(1)} \text{ and } (x_\alpha^{(2)}, y_\beta^{(2)}) \text{ is } (1,0) \end{cases}$$

and U by (B 1), then the statistic $(n_1 n_2)^{-1}W$ is the same as U.

Now $E\{t(X_\alpha, Y_\beta)\}$ is well defined and $E\{t(X_\alpha, Y_\beta)\}^2 \leqslant 1$ under null and alternative hypotheses. Hence as $n_1 \to \infty$ with $\lim n_1/n_2$ fixed and non-zero, the distribution of U is asymptotically normal. We have shown

$$\frac{W}{\sqrt{\{\mathrm{var}(W|H_0)\}}} \text{ is asymptotically } N(0,1) \tag{B 2}$$

and we wish to show

$$\frac{W}{\sqrt{\{\mathrm{var}(W|P,H_0)\}}} \text{ is asymptotically } N(0,1). \tag{B 3}$$

Now if

$$p - \lim \frac{n_1^{-3}\,\mathrm{var}(W|P,H_0)}{n_1^{-3}\,\mathrm{var}(W|H_0)} \to 1$$

as $n_1 \to \infty$ and $\lim n_1/n_2$ exists we obtain (B 2) from a convergence theorem of Cramér (1946, p. 254).

But

$$\mathrm{var}(W|H_0) = E_P\,\mathrm{var}(W|P,H_0) + \mathrm{var}_P E(W|P,H_0),$$

where the expectation is over all possible patterns that could arise.

Under the null hypothesis, the number of individuals failing and being censored at the $2s$ points in the general pattern can be considered as an outcome in multinomial sampling. The sample size is n_1+n_2 and the sum of probabilities over the $2s$ points is one.

Now $E(W|P,H_0) = 0$, so we need to consider

$$\{n_1^{-3}\,\mathrm{var}(W|P,H_0)\}/\{n_1^{-3}E_P\,\mathrm{var}(W|P,H_0)\} \quad \text{as} \quad n_1 \to \infty.$$

The numerator is a polynomial function of (m_i, M_i, l_i, L_i) and, by a proposition quoted by Cramér (1946, p. 255), converges in probability to the constant obtained by replacing the above variables by

their expectations from the multinomial. The denominator is the average of the numerator over all possible patterns and approaches the same constant (to $O(1/n_1)$). Hence, we obtain (B 3) and (B 2) follows from the convergence theorem of Cramér. The result holds for patterns that arise randomly in the manner described.

APPENDIX C

The definition of asymptotic efficiency of W relative to F is given by (8·1) and we now proceed to evaluate the various terms for case (a) and (b).

Case (a)

For the F test, we have

$$z = \tfrac{1}{2}\log(\bar{t}_1/\bar{t}_2)$$

and we wish to find

$$E(z) = E_s E(z|s),$$
$$\operatorname{var}(z|H_0) = E_s \operatorname{var}(z|H_0, s) + \operatorname{var}_s E(z|H_0, s).$$

Here the pattern of observations is defined by the total sample size $(2n)$ and the number of failure observations (s) prior to T. We consider expectations and variances in the conditional universe where $s = 2n - r_1 - r_2$ is fixed, and then allow variations in s. The calculations will be asymptotic as $n, s \to \infty$. Under H_0, s has a binomial distribution with $E(s) = 2n(1 - e^{-T\phi})$.

Because $E(\bar{t}_1) = 1/\phi$ and $E(\bar{t}_2) = 1/(\phi\theta)$, we find

$$E(z) = E_s E(z|s) \cong \tfrac{1}{2}\log\theta, \tag{C 1}$$
$$\left.\frac{\partial E(z)}{\partial\theta}\right|_{\theta=1} = \frac{1}{2}.$$

Also,

$$\operatorname{var}(z|H_0) \cong E_s \frac{1}{2}\left(\frac{1}{2(\frac{1}{2}s)} + \frac{1}{2(\frac{1}{2}s)}\right) \tag{C 2}$$
$$\cong \frac{1}{2n(1 - e^{-T\phi})}.$$

For the W test, we have $W = \sum_{i,j} U_{ij}$ as defined by (3·1). Now

$$E(W) = n^2\{\Pr(X_i > Y_j) + \Pr(X_i' \geqslant Y_j) - \Pr(X_i < Y_j) - \Pr(X_i \leqslant Y_j')\}, \tag{C 3}$$

where X_i, X_i' are random variables of times to failure, censoring determined by $f_1(x)$ and similarly Y_j, Y_j' are determined by $f_2(y)$. Here, $X_i' \equiv Y_j' \equiv T$ and the probability of being censored at T is $e^{-T\phi}, e^{-T\phi\theta}$, respectively. The probabilities are obtained as follows:

$$\Pr(X_i > Y_j) + \Pr(X_i' \geqslant Y_j) = \int_0^T \theta\phi\, e^{-\theta\phi u} e^{-\phi u}\, du$$
$$= \frac{\theta}{\theta+1}(1 - e^{-T\phi(\theta+1)})$$

and

$$\Pr(X_i < Y_j) + \Pr(X_i \leqslant Y_j') = \int_0^T \phi\, e^{-\phi u} e^{-\theta\phi u}\, du$$
$$= \frac{1}{\theta+1}(1 - e^{-T\phi(\theta+1)}).$$

Hence,

$$E(n^{-2}W) = \frac{(\theta-1)}{(\theta+1)}(1 - e^{-T\phi(\theta+1)}) \tag{C 4}$$

and

$$\left.\frac{\partial E(n^{-2}W)}{\partial\theta}\right|_{\theta=1} = \tfrac{1}{2}(1 - e^{-2T\phi}).$$

Now

$$\operatorname{var}(n^{-2}W|H_0) = n^{-4}E\{\sum_{i,j} U_{ij} - E(\sum_{i,j} U_{ij})|H_0\}^2$$
$$= n^{-4}E\{\sum_{i,j} U_{ij}^2 + \sum_{i \neq i'} U_{ij}U_{i'j} + \sum_{j \neq j'} U_{ij}U_{ij'} + \sum_{\substack{i \neq i' \\ j \neq j'}} U_{ij}U_{i'j'}|H_0\},$$

since $E(\sum_{i,j} U_{ij}|H_0) = 0$. To evaluate the four terms necessary for the variance, note that there are only n^2 terms of the type $E(U_{ij}^2)$ so that the total contribution of the first term is $O(n^{-2})$. Then

$$E(\sum_{i \neq i'} U_{ij}U_{i'j}|H_0) = E(\sum_{j \neq j'} U_{ij}U_{ij'}|H_0),$$

by symmetry and there are $n^2(n-1)$ terms in each summation. Also $E(U_{ij}U_{ij'}|H_0) = 0$ since U_{ij} is independent of $U_{ij'}$ and each has expectation zero. Thus

$$\operatorname{var}(n^{-2}W|H_0) \cong 2n^{-1}E(U_{ij}U_{ij'}|H_0).$$

Using the conditions in (3·1), we have

$$\begin{aligned} U_{ij}U_{ij'} = 0 \quad \text{otherwise,} \\ &-1(X_i < Y_j, X_i > Y_{j'}), (X_i > Y_j, X_i < Y_{j'}), \\ &(X_i \leqslant Y'_j, X_i > Y_{j'}), (X_i > Y_j, X_i \leqslant Y'_{j'}), \\ &+1(X_i > Y_j, Y_{j'}), (X_i < Y_j, Y_{j'}), \\ &(X_i < Y_j, Y'_{j'}), (X_i < Y'_j, Y_{j'}), \\ &(X'_i > Y_j, Y_{j'}), (X_i < Y'_j, Y'_{j'}). \end{aligned}$$

We now wish to calculate the probabilities of the various events on the right-hand side. Let p_d be the probability of failure under H_0 and p_s be the probability of a censored observation. Also, let $f_d(x)$, $f_s(x')$ be conditional probability density functions of time to failure, censoring, respectively. For case (a), these are all simple to write down.

Thus,

$$p_d = 1 - e^{-T\phi}, \tag{C 5}$$

$$p_s = e^{-T\phi} \tag{C 6}$$

and

$$f_d(x) = \frac{\phi\, e^{-\phi x}}{p_d} \quad (0 \leqslant x < T),$$

$$f_s(x') = \frac{e^{-T\phi}}{p_s} = 1 \quad (x' = T).$$

For case (a), $f_s(x')$ is a discrete probability, but the notation is retained to be analogous with case (b).

Now under H_0,

$$\Pr(X_i > Y_j, Y_{j'}) = \Pr(X_i < Y_j, Y_{j'}) = \tfrac{1}{3}$$

$$\Pr(X_i < Y_j, X_i > Y_{j'}) = \Pr(X_i > Y_j, X_i < Y_{j'}) = \tfrac{1}{6}.$$

Thus

$$\begin{aligned} E(U_{ij}U_{ij'}|H_0) = {} & p_d^3(\tfrac{1}{3} + \tfrac{1}{3}) + 2p_d^2 p_s \Pr\{X_i < \min(Y_j, Y'_{j'})\} \\ & + p_d^2 p_s \Pr\{X'_i > \max(Y_j, Y_{j'})\} \\ & + p_s^2 p_d \Pr\{X_i < \min(Y'_j, Y'_{j'})\} \\ & - p_d^3(\tfrac{1}{6} + \tfrac{1}{6}) - 2p_d^2 p_s \Pr\{X_i < Y'_j, X_i > Y_{j'}\}, \end{aligned}$$

where the time to failure variables follow $f_d(x)$ and the time to censoring variables follow $f_s(x')$.

Now $\Pr\{X'_i > \max(Y_j, Y_{j'})\} = 2\Pr\{X_i < Y'_j, X_i > Y_{j'}\}$. Hence

$$E(U_{ij}U_{ij'}|H_0) = \tfrac{1}{3}p_d^3 + 2p_d^2 p_s \Pr\{X_i < \min(Y_j, Y'_{j'})\} + 2p_s^2 p_d \Pr\{X_i < \min(Y'_j, Y'_{j'})\}, \tag{C 7}$$

and the probabilities can be written down immediately:

$$\Pr\{X_i < \min(Y_j, Y'_{j'})\} = \tfrac{1}{2}, \quad \Pr\{X_i < \min(Y'_j, Y'_{j'})\} = 1.$$

Thus

$$E(U_{ij}U_{ij'}|H_0) = \tfrac{1}{3}(1 - e^{-T\phi})^3 + e^{-T\phi}(1 - e^{-T\phi})$$

and

$$\operatorname{var}(n^{-2}W|H_0) = n^{-1}\{\tfrac{2}{3}(1 - e^{-T\phi})^3 + 2e^{-T\phi}(1 - e^{-T\phi})\}. \tag{C 8}$$

Finally, the A.R.E. of W to F is obtained by substituting (C 1), (C 2), (C 4) and (C 8) into (8·1) to get

$$\text{A.R.E.} = \frac{(1 - e^{-2T\phi})^2}{\tfrac{4}{3}(1 - e^{-T\phi})^4 + 4e^{-T\phi}(1 - e^{-T\phi})^2}.$$

Values of A.R.E. for case (a) are given in Table 1 for various $T\phi$.

Case (b)

In this situation, $2n$ patients are entered into study according to a uniform distribution in the fixed interval 0 to T and fail according to an exponential distribution. In the group receiving treatment A, the probability of a patient entering in any interval of time (Δt) is $(\Delta t)/T$ and failure is according to $f_1(x)$. We have

$$\begin{aligned} \Pr(\text{patient fails at 'age' } x) &= \int_0^{T-x} \frac{\phi}{T} e^{-\phi x}\, dx\, dv \\ &= \frac{(T-x)}{T}\, \phi\, e^{-\phi x}\, dx \quad (0 \leqslant x \leqslant T), \end{aligned} \tag{C 9}$$

$$\Pr(\text{patient censored at 'age' } x') = \frac{1}{T} e^{-\phi x'}\, dx' \quad (0 \leqslant x' \leqslant T) \tag{C 10}$$

where 'age' is measured from time of entry into study.

Now
$$p_d = \int_0^T \frac{(T-u)}{T}\,\phi\, e^{-\phi u}\,du = 1-\frac{1}{T\phi}(1-e^{-T\phi}), \tag{C 11}$$
$$p_s = \int_0^T \frac{1}{T} e^{-\phi u}\,du = \frac{1}{T\phi}(1-e^{-T\phi}) \tag{C 12}$$

and $f_d(x), f_s(x')$ are the probability density parts of (C 9), (C 10) divided by p_d, p_s, respectively. Similarly, $f_d(y)$ and $f_s(y')$ are defined by replacing ϕ by $\theta\phi$ in p_d, p_s, $f_d(x)$ and $f_s(x')$.

For the F test, we transform l_2/l_1 to z as before and we have
$$\frac{\partial E(z)}{\partial\theta} = -\frac{1}{2E(l_2)}\frac{\partial E(l_2)}{\partial\theta}. \tag{C 13}$$

Now
$$E(l_2|n-r_2) = \frac{1}{n-r_2}\{(n-r_2)E(Y)+r_2E(Y')\},$$

where Y, Y' follow $f_d(y)$, $f_s(y')$, respectively. Then, $E(Y)$ and $E(Y')$ are easily evaluated and when substituted in (C 13) we find $E(l_2|n-r_2) = 1/(\theta\phi)$ and so
$$\left.\frac{\partial E(z)}{\partial\theta}\right|_{\theta=1} = \frac{1}{2}. \tag{C 14}$$

Under H_0, the expected number of individuals failing in the two groups before T is $2np_d$, so that
$$\operatorname{var}(z|H_0) \cong \frac{1}{2n\{1-1/T\phi(1-e^{-T\phi})\}}. \tag{C 15}$$

For the W test, $E(W)$ is defined by (C 3) and the probabilities needed are found using (C 9) and (C 10):
$$\begin{aligned}\Pr(X_i > Y_j) &= \iint_{u>v} \frac{(T-u)}{T}\,\phi\, e^{-\phi u}\,\frac{(T-v)}{T}\,\theta\phi\, e^{-\theta\phi v}\,du\,dv \quad (0 \leqslant u, v < T)\\ &= \frac{\theta}{\theta+1} - \frac{1}{T\phi(\theta+1)^2}\{\theta(\theta+3)-e^{-T\phi}(\theta+1)^2\}\\ &\quad + \frac{1}{[T\phi(\theta+1)]^2\theta(\theta+1)}\{\theta^2(\theta+3)-e^{-T\phi}(\theta+1)^3+e^{-T\phi(\theta+1)}(3\theta+1)\},\end{aligned} \tag{C 16}$$
$$\begin{aligned}\Pr(X_i' > Y_j) &= \iint_{u>v} \frac{1}{T}\,e^{-\phi u}\,\frac{(T-v)}{T}\,\theta\phi\, e^{-\theta\phi v}\,du\,dv \quad (0 \leqslant u, v \leqslant T)\\ &= \frac{\theta}{T\phi(\theta+1)} - \frac{e^{-T\phi}}{T\phi} - \frac{1}{[T\phi(\theta+1)]^2}\left\{\theta - \frac{e^{-T\phi}(\theta+1)^2}{\theta} + \frac{(2\theta+1)}{\theta}\,e^{-T\phi(\theta+1)}\right\}.\end{aligned} \tag{C 17}$$

The $\Pr(X_i < Y_j)$ and $\Pr(X_i < Y_j')$ are obtained by replacing θ by $1/\theta$ and ϕ by $\theta\phi$ in (C 16), (C 17), respectively. Substituting these results in (C 3), we have
$$\left.\frac{\partial E(n^{-2}W)}{\partial\theta}\right|_{\theta=1} = \frac{1}{2}\left\{1-\frac{1}{T\phi}+\frac{1}{2(T\phi)^2}(1-e^{-2T\phi})\right\}. \tag{C 18}$$

The value of $\operatorname{var}(n^{-2}W|H_0)$ is found in exactly the same way as before, with p_d and p_s of (C 11) and (C 12) replacing (C 5) and (C 6) in the equation for $E(U_{ij}U_{ij'}|H_0)$ given by (C 7). We now need to evaluate $\Pr\{X_i < \min(Y_j, Y_{j'}')\}$ and $\Pr\{X_i < \min(Y_j', Y_{j'}')\}$.

Now under H_0,
$$\begin{aligned}\Pr(Y_j < u) &= \int_0^u \frac{(T-v)\phi}{Tp_d}\,e^{-\phi v}\,dv\\ &= \frac{1}{p_d}\left\{(1-e^{-\phi u})-\frac{1}{T\phi}(1-e^{-\phi u}[1+\phi u])\right\}\end{aligned}$$

and
$$\begin{aligned}\Pr(Y_j' < u) &= \int_0^u \frac{1}{Tp_s}\,e^{-\phi v}\,dv\\ &= \frac{1}{T\phi p_s}(1-e^{-\phi u}).\end{aligned}$$

Therefore
$$\begin{aligned}\Pr\{X_i < \min(Y_j, Y_{j'}')\} &= \int_0^T \frac{(T-u)}{Tp_d}\,\phi\, e^{-\phi u}\left\{1-\frac{1}{T\phi p_s}(1-e^{-\phi u})\right\}\\ &\quad\times\left\{1-\frac{1}{p_d}\left[(1-e^{-\phi u})-\frac{1}{T\phi}(1-e^{-\phi u}[1+\phi u])\right]\right\}du\end{aligned}$$

and

$$\Pr\{X_i < \min(Y'_j, Y'_{j'})\} = \int_0^T \frac{(T-u)}{Tp_d}\phi\left\{1-\frac{1}{T\phi p_s}(1-e^{-\phi u})\right\}^2 du;$$

these integrals are simple, though somewhat laborious, to evaluate. The results are

$$\Pr\{X_i < \min(Y_j, Y'_{j'})\} = \frac{1}{p_d^2 p_s}\left\{\frac{1}{T\phi}\left(\tfrac{1}{3}-\tfrac{1}{2}e^{-T\phi}\right)+\frac{1}{(T\phi)^2}\left(-\tfrac{5}{9}+\tfrac{3}{4}e^{-T\phi}-e^{-2T\phi}\right)\right.$$
$$\left.+\frac{1}{(T\phi)^3}\left(\tfrac{5}{27}-\tfrac{3}{4}e^{-T\phi}+e^{-2T\phi}-\tfrac{47}{108}e^{-3T\phi}\right)\right\}$$

and

$$\Pr\{X_i < \min(Y'_j, Y'_{j'})\} = \frac{1}{p_d p_s^2}\left\{\frac{1}{(T\phi)^2}\left(\tfrac{1}{3}-e^{-T\phi}+e^{-2T\phi}\right)+\frac{1}{(T\phi)^3}\left(-\tfrac{1}{9}+\tfrac{1}{2}e^{-T\phi}-e^{-2T\phi}+\tfrac{11}{18}e^{-3T\phi}\right)\right\}.$$

With these probabilities, we can now evaluate $E(U_{ij}U_{ij'}|H_0)$ and $\operatorname{var}(n^{-2}W|H_0)$. We have

$$\operatorname{var}(n^{-2}W|H_0) \cong n^{-1}\left\{\frac{2}{3}-\frac{2}{3(T\phi)}+\frac{4}{9(T\phi)^2}-\frac{4}{27(T\phi)^3}(1-e^{-3T\phi})\right\}. \tag{C 19}$$

Substituting (C 14), (C 15), (C 18) and (C 19) into (8·1) we can calculate the asymptotic efficiency of W relative to F. This is done for various values of $T\phi$ in Table 1.