Self-consistent: p. 840
Re-distribute to the right: p. 842

THE TWO SAMPLE PROBLEM WITH CENSORED DATA

BRADLEY EFRON
STANFORD UNIVERSITY

1. Introduction

A medical investigator attempting to compare two different treatments for, say, prolongation of life among disease victims, often finds himself in the following situation: at time T, when it is necessary to end the experiment, or at least evaluate the results up to that time, a certain number of the patients in each treatment group will still be alive. His data will then be represented by two sets of numbers which might look like $x_1, x_2, x_3+, x_4, x_5+, x_6, \cdots, x_m$ and $y_1, y_2+, y_3+, y_4, \cdots, y_n$. Here x_1 and x_2 would represent actual lifetimes, while x_3+, a "censored" observation, represents a lifetime known only to exceed x_3. If all the patients in both treatment groups were treated at time 0, then every $+$ value would be equal to T, a situation that has been investigated by Halperin [1]. Frequently, however, patients enter the investigation at different times after it has begun, and the $x+$ and $y+$ values may range from 0 to T. Such a situation, of course, complicates the comparison of the two treatments, particularly if the mechanism censoring the x values is different from that censoring the y values. This may happen, for instance, if the x sequence was run some time ago, so that nearly all the patients have been observed to their death times, while the y sequence is begun later, and contains many censored observations.

Gehan [2] and Gilbert [3] have independently proposed the same extension of the Wilcoxon statistic as a solution to the two sample problem with censored data. In this paper the problem is discussed further, and a different test statistic is proposed, which is shown to be, in some ways, superior to the Gehan-Gilbert statistic.

2. A statement of the problem and some notation

Suppose $x_1^0, x_2^0, \cdots, x_m^0$ are independent, identically distributed random variables, having $F^0(s) = P\{x_i^0 \geqq s\}$ as their common right sided cumulative distribution function (c. d. f.). (Because of the censorship from the right, this is a more convenient function to deal with than the usual left c. d. f. Note that $F^0(s)$ is a left continuous, nonincreasing function of s, and that $F^0(-\infty) = 1$, $F(\infty) = 0$.) Likewise, let $y_1^0, y_2^0, \cdots, y_n^0$ be independent, identically distributed

random variables with common right c. d. f., $G^0(s) = P\{y_j^0 \geqq s\}$. Both F^0 and G^0 will be assumed continuous. Our null hypothesis, which we wish to test, is

$$H_0: \quad F^0 = G^0, \tag{2.1}$$

that is, the x_i^0 and y_j^0 random variables have the same distribution. What we are considering as alternatives to H_0 will become apparent as we discuss the various test statistics.

Unfortunately, we are not allowed to observe the x_i^0 and y_j^0 values directly. Rather, we are given two additional sequences of numbers, $u_1, u_2, \cdots, u_m$, and $v_1, v_2, \cdots, v_n$, and our observations consist of the minima,

$$\begin{aligned} x_1 &= \min(x_1^0, u_1), \quad x_2 = \min(x_2^0, u_2), \quad \cdots, x_m = \min(x_m^0, u_m), \\ y_1 &= \min(y_1^0, v_1), \quad y_2 = \min(y_2^0, v_2), \quad \cdots, y_n = \min(y_n^0, v_n). \end{aligned} \tag{2.2}$$

In addition, we know which of the x_i are actually x_i^0, that is, uncensored observations, and which are the u_i, and likewise for the y_j. This information will be denoted by the two (random) sequences, $\delta_1, \delta_2, \cdots, \delta_m$ and $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$, where

$$\begin{aligned} \delta_i &= \begin{cases} 1 \text{ if } x_i = x_i^0 \\ 0 \text{ if } x_i < x_i^0, \end{cases} \qquad (\text{so } x_i = u_i); \\ \epsilon_0 &= \begin{cases} 1 \text{ if } y_j = y_j^0 \\ 0 \text{ if } y_j < y_j^0, \end{cases} \qquad (\text{so } y_j = v_j). \end{aligned} \tag{2.3}$$

For the purposes of computing means, variances, efficiencies, and so forth, it is convenient to assume that the u_i and v_j are independent random variables themselves, with continuous right c. d. f.'s

$$\begin{aligned} H(s) &= P\{u_i \geqq s\}, & i &= 1, 2, \cdots, m, \\ I(s) &= P\{v_j \geqq s\}, & j &= 1, 2, \cdots, n, \end{aligned} \tag{2.4}$$

respectively. (However, the reader should keep in mind that such assumptions are not essential for the application of the various tests. This point is discussed further in section 5.) Under these assumptions, the x_i and the y_j are mutually independent random variables with right c. d. f.'s

$$\begin{aligned} F(s) &= F^0(s)H(s) = P\{x_i \geqq s\}, & i &= 1, 2, \cdots, m, \\ G(s) &= G^0(s)I(s) = P\{y_j \geqq s\}, & j &= 1, 2, \cdots, n, \end{aligned} \tag{2.5}$$

respectively. The δ_i and ϵ_j are mutually independent Bernoulli random variables, with

$$\begin{aligned} P\{\delta_i = 1\} &= P\{H \geqq F^0\} = -\int_{-\infty}^{\infty} H(s)dF^0(s), \\ P\{\epsilon_j = 1\} &= P\{I \geqq G^0\} = -\int_{-\infty}^{\infty} I(s)dG^0(s). \end{aligned} \tag{2.6}$$

The minus sign compensates for F^0 and G^0 being decreasing functions of s. Throughout this paper, such notation as $P\{F \geqq G\}$ will mean "the probability that a random variable with right c. d. f. $F(s)$ is greater than an independent random variable with right c. d. f. $G(s)$," that is, $P\{F \geqq G\} = -\int_{-\infty}^{\infty} F(s)dG(s)$.

In general, the random variable x_i will *not* be independent of δ_i, and likewise for y_j and ϵ_j. The example in section 9 is especially simple because independence does hold in that special case.

3. The test statistic of Gehan and Gilbert

Let us define a "scoring function"

$$Q_G(x_i, y_j, \delta_i, \epsilon_j) = \begin{cases} 1 \text{ if } x_i \geqq y_j, & \epsilon_j = 1, \\ 0 \text{ if } x_i < y_j, & \delta_i = 1, \\ 1/2 \text{ otherwise,} \end{cases} \tag{3.1}$$

and a statistic

$$W_G = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} Q_G(x_i, y_j, \delta_i, \epsilon_j). \tag{3.2}$$

Because of the continuity of F, G, H, and I, the condition $x_i \geqq y_j$ could be replaced with $x_i > y_j$. However, throughout this paper, the definitions have been chosen to be consistent even when the c. d. f.'s are not continuous.

The statistic W_G has been proposed independently by Gehan [2] and Gilbert [3] (whose names, fortunately, both begin with G) as a reasonable extension of Wilcoxon's statistic to the case of censored data. It is easily seen that mnW_G equals the number of (x_i, y_j) pairs where x_i^0 is known definitely to be larger (or as large as) y_j^0, plus one half the number of pairs where, on the basis of the given data, x_i^0 may be larger or may be smaller than y_j^0. It is instructive to follow Gilbert and rewrite Q_G as

$$Q_G(x_i, y_j, \delta_i, \epsilon_j) = \begin{cases} 1 \text{ if } y_j^0 \leqq \min(x_i^0, u_i, v_j) \\ 0 \text{ if } x_i^0 < \min(y_j^0, v_j, u_i) \\ 1/2 \text{ if } u_i < \min(x_i^0, y_j^0, v_j) \\ 1/2 \text{ if } v_j < \min(x_i^0, y_j^0, u_i) \end{cases} \tag{3.3}$$

which a simple enumeration of cases shows is equivalent to (3.1). This yields the expectation of W_G, namely

$$\begin{aligned} EW_G &= EQ_G(x_i, y_j, \delta_i, \epsilon_j) \\ &= P\{F^0HI \geqq G^0\} + \frac{1}{2}[P\{F^0G^0I > H\} + P\{F^0G^0H > I\}], \end{aligned} \tag{3.4}$$

where F^0HI is the right c. d. f. of $\min(x_i^0, u_i, v_j)$, and so forth. Under the null hypothesis, H_0: $F^0 = G^0$, we have $P\{F^0HI \geqq G^0\} = P\{G^0HI \geqq F^0\}$, and

$$\begin{aligned} EW_G = \frac{1}{2}[&P\{F^0HI \geqq G^0\} + P\{G^0HI \geqq F^0\} \\ &+ P\{F^0G^0I > H\} + P\{F^0G^0H > I\}] = \frac{1}{2}, \end{aligned} \tag{3.5}$$

independent of the distributions H and I. (This fact was first brought to my attention by Dr. Nathan Mantel, who was kind enough to send me a preprint of his paper [4].)

Gilbert gives the following formula for the variance of W_G under H_0

$$\text{Var}_{H_0}(W_G) = \frac{1}{12mn}\,[3P\{HI > (F^0)^2\} + (n-1)P\{HI^2 > (F^0)^3\} + (m-1)P\{H^2I > (F^0)^3\}]. \tag{3.6}$$

In what follows, some advantages and disadvantages of the W_G test statistic will be discussed, particularly in comparison with the $\hat{W}$ test introduced in section 8.

4. Asymptotically nonparametric tests

The statistic W_G is not nonparametric, since even under H_0, its distribution will depend on the relationship between H, I, and $F^0 = G^0$ (this can be seen in (3.6), for instance). However, it is "asymptotically nonparametric" in the following sense: if both m and n go to ∞ in such a way that $\lim[m/(m+n)] = \lambda$, with $0 < \lambda < 1$, then, under H_0,

$$(m+n)^{1/2}\left(W_G - \frac{1}{2}\right) \underset{\text{law}}{\to} N\left[0, \frac{1}{12}\left(\frac{1}{\lambda}\,\sigma_1^2 + \frac{1}{1-\lambda}\,\sigma_2^2\right)\right], \tag{4.1}$$

where

$$\begin{aligned} \sigma_1^2 &= P\{HI^2 > (F^0)^3\} \\ \sigma_2^2 &= P\{H^2I > (F^0)^3\}. \end{aligned} \tag{4.2}$$

Here $N(\mu, \sigma^2)$ represents the normal law with mean μ and variance σ^2. (This is a weaker than necessary consequence of the two sample U statistic theorem [5].) Moreover, σ_1^2 and σ_2^2 have consistent estimators as m and n go to infinity. Letting, $\hat{H}$, $\hat{I}$, and $\hat{F}$ be the usual (right sided) sample c. d. f.'s calculated from the u_i, v_j and x_i values, respectively, then

$$\hat{\sigma}_1^2 = -\int_0^{\hat{x}} \hat{H}(s)\hat{I}^2(s)\, d\left(\frac{\hat{F}(s)}{\hat{H}(s)}\right)^2 \tag{4.3}$$

will estimate σ_1^2 consistently, where $\hat{x} = \max\{x: \hat{H}(x) > 0\}$, and likewise for $\hat{\sigma}_2^2$.

Any asymptotically nonparametric statistic can be used to construct an asymptotically level α test of H_0. The only statistics considered in this paper will be those having the asymptotically nonparametric property. It must be remembered that in actual practice, the consistent estimator of the variance necessary to carry out the level α significance test may be more or less difficult to obtain, and this is a definite factor in comparing different test statistics.

The following well known result from parametric theory offers some insight into the asymptotically nonparametric property, and will be useful in its own right in section 9. Suppose $f_\theta(x)$ is a density function depending on a $k+1$ dimensional parameter $\theta = (\theta^0, \theta^1, \theta^2, \cdots, \theta^k)$, and having information matrix $I_\theta = \{I_{ij} | 0 \leqq i, j \leqq k\}$,

$$I_{ij} = E_\theta \frac{\partial \log f_\theta(x)}{\partial \theta^i} \frac{\partial \log f_\theta(x)}{\partial \theta^j}. \tag{4.4}$$

It is desired to test H_0: $\theta^0 = \theta_0^0$ on the basis of repeated independent observations $x_1, x_2, \cdots, x_n, \cdots$ from $f_\theta(x)$. Let C be the class of all test statistics $t_n(x_1, x_2, \cdots, x_n)$, such that under H_0 we have

$$n^{1/2}(t_n - \mu) \underset{\text{law}}{\to} N(0, \sigma^2(\theta_0)) \tag{4.5}$$

for some constant μ not depending on $\theta_0 \in H_0$. Then, under suitable regularity conditions, the efficacy of any member of C is bounded above by $1/I^{00}$, where I^{00} is the upper left entry of $I_{\theta_0}^{-1}$. That is,

$$\lim_{n\to\infty} \frac{\left(\left.\frac{\partial Et_n}{\partial\theta^0}\right|_{\theta=\theta_0}\right)^2}{n \operatorname{Var}_{\theta_0}(t_n)} \leqq \frac{1}{I^{00}}, \tag{4.6}$$

for all $t_n \in C$. Moreover, the maximum likelihood estimate $\hat{\theta}^0$ attains the upper bound.

The bound above can be obtained formally from the multivariate Cramér-Rao inequality. A discussion of this type of theorem, with the suitable conditions, can be found in [6]. The class C, which corresponds to our asymptotically nonparametric tests, is essential to the theorem. Usually there will exist statistics outside of C that do not satisfy (4.6).

A genuinely nonparametric test for H_0 is mentioned in section 5, but is shown to be very inefficient. Both Gehan and Gilbert propose the same nonparametric test for H_0, under the additional assumption that $H = I$, that is, identical censoring distributions (namely, the permutation test based on W_G). However, this test is not even asymptotically nonparametric when $H \neq I$. It seems doubtful that a reasonably efficient nonparametric test for H_0 could be constructed in the general case, but the question remains open.

5. A test of H_0 based on cross censorship

It is not difficult to generate whole families of asymptotically nonparametric tests of the hypothesis H_0: $F^0 = G^0$ as an extension of Gehan and Gilbert's method. For example, let $t(x, y)$ be any bounded real valued function, and define the scoring function $Q_t(x_i, y_j, \delta_i, \epsilon_j)$ as

$$Q_t(x_i, y_j, \delta_i, \epsilon_j) = \begin{cases} t(x_i, y_j) & \text{if } x_i \geqq y_j \text{ and } \epsilon_j = 1, \\ -t(y_j, x_i) & \text{if } x_i < y_j \text{ and } \delta_i = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

Then under H_0 the statistic

$$W_t = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} Q_t(x_i, y_j, \delta_i, \epsilon_j) \tag{5.2}$$

will have expectation zero, since

$$\begin{aligned} EW_t &= EQ_t(x_i, y_j, \delta_i, \epsilon_j) \\ &= E[t(x_i^0, y_j^0) | y_j^0 \leqq \min(x_i^0, u_i, v_j)] - E[t(y_j^0, x_i^0) | x_i^0 < \min(y_j^0, u_i, v_j)] \\ &= 0 \end{aligned} \tag{5.3}$$

under H_0 by symmetry. (If we let $t(x, y) \equiv 1$, then $2W_t - 1 = W_G$.) Asymptotic normality and the existence of a consistent estimator of the large sample variance follows again from the two sample U statistic theorem.

In this section, a somewhat different type of asymptotically nonparametric test statistic will be discussed briefly. Though the specific statistic derived will be shown to be inefficient relative to both W_G and the $\hat{W}$ statistic introduced in section 8, the method of its construction is simple, and applicable in a wide variety of situations. In the sequel, it will also provide several points of comparison with the $\hat{W}$ statistic.

Suppose, for a moment, that in addition to the data x_i, δ_i, y_j, ϵ_j, $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, n$, we are also given the c. d. f.'s of the two censoring distributions, H and I. This is sometimes a realistic assumption. For instance, if the entry of patients into the experiment is random in time, then H and I may be assumed to be uniform over the period of observation. Using a table of random numbers, draw two new sets of mutually independent random variables, say

$$\begin{aligned} &u_1^*, u_2^*, \cdots, u_n^* \sim H \\ &v_1^*, v_2^*, \cdots, v_m^* \sim I, \end{aligned} \tag{5.4}$$

and define the "cross censored" observations

$$\begin{aligned} &x_1^* = \min(x_1, v_1^*), \quad x_2^* = \min(x_2, v_2^*), \quad \cdots, x_m^* = \min(x_m, v_m^*), \\ &y_1^* = \min(y_1, u_1^*), \quad y_2^* = \min(y_2, u_2^*), \quad \cdots, y_n^* = \min(y_n, u_n^*). \end{aligned} \tag{5.5}$$

The x_i^* and y_j^* are mutually independent random variables, having right c. d. f.'s F^0HI and G^0HI, respectively, since $x_i^* = \min(x_i^0, u_i, v_i^*)$ and $y_j^* = \min(y_j^0, v_j, u_j^*)$. Under the null hypothesis H_0, $F^0HI = G^0HI$, and the usual Wilcoxon test is applicable. That is, define

$$\begin{aligned} Q(x_i^*, y_j^*) &= \begin{cases} 1 \text{ if } x_i^* \geqq y_j^*, \\ 0 \text{ if } x_i^* < y_j^*; \end{cases} \\ W &= \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} Q(x_i^*, y_j^*). \end{aligned} \tag{5.6}$$

Under H_0: $F^0 = G^0$, this W will have the usual Wilcoxon distribution with

$$EW = \frac{1}{2}, \qquad \operatorname{Var}_{H_0} W = \frac{1}{12} \frac{m + n + 1}{mn}. \tag{5.7}$$

When $F^0 \neq G^0$, the expectation is given by

$$EW = P\{F^0HI \geqq G^0HI\}. \tag{5.8}$$

Since we have used a table of random numbers in the construction of W, we know that it is inefficient. Using the usual Rao-Blackwell method, we can get an improved statistic by considering

$$W_c = E(W|\text{information}), \tag{5.9}$$

the conditional expectation being taken with respect to all of our available information: the x_i, y_j, δ_i, ϵ_j, and the functions H and I. We will then have

$$EW_c = EW = P\{F^0III \geqq G^0III\},$$
$$\text{Var } W_c \leqq \text{Var } W, \tag{5.10}$$

for all choices of F^0, G^0, H, and I.

It is not difficult to express W_c in an easily computable form. Let F and G be the usual (right) sample c. d. f.'s of the x_i and y_j observations respectively; that is, define

$$N_x(s) = (\text{number of } x_i \geqq s), \qquad N_y(s) = (\text{number of } y_j \geqq s),$$
$$\hat{F}(s) = \frac{N_x(s)}{m}, \qquad \hat{G}(s) = \frac{N_y(s)}{n}. \tag{5.11}$$

(Note that $\hat{F}$ represents a distribution with mass $1/m$ at the points $x_1, x_2, \cdots, x_m$ and is an estimate of F, *not* F^0. Likewise, $\hat{G}$ puts mass $1/n$ at $y_1, y_2, \cdots, y_n$ and estimates G, not G^0.)

Let $\hat{x}$, $\hat{y}$, u^*, and v^* be independent random variables with right c. d. f.'s $\hat{F}$, $\hat{G}$, H, and I, respectively, and define $\hat{x}^* = \min(\hat{x}, v^*)$, $\hat{y}^* = \min(\hat{y}, u^*)$. Then

$$\begin{aligned} &P\{\hat{x}^* \geqq \hat{y}^*\} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} P\{\hat{x}^* \geqq \hat{y}^* | \hat{x} = x_i, \hat{y} = y_j\} P\{\hat{x} = x_i, \hat{y} = y_j\} \\ &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} P\{x_i^* \geqq y_j^* | x_i, y_j\} \\ &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} E[Q(x_i^*, y_j^*) | x_i, y_j] \\ &= E(W|\text{information}) = W_c. \end{aligned} \tag{5.12}$$

Since by definition $P\{\hat{x}^* \geqq \hat{y}^*\} = P\{\hat{F}I \geqq \hat{G}H\}$, we have

$$\begin{aligned} W_c &= P\{\hat{F}I \geqq \hat{G}H\} \\ &= -\int_{-\infty}^{\infty} \hat{F}(s)I(s)\, d[\hat{G}(s)H(s)], \end{aligned} \tag{5.13}$$

in close analogy with the integral expression for the usual Wilcoxon statistic in the uncensored situation, which is, in our type of notation, $-\int_{-\infty}^{\infty} \hat{F}(s)\, d\hat{G}(s)$.

It is easy to show (by the methods used in section 8, or by the theorem on U statistics) that as we let m and n go to infinity so that $\lim[m/(m+n)] = \lambda$, where $0 < \lambda < 1$, then W_c is asymptotically normal, and under H_0,

$$(m+n)^{1/2}\left(W_c - \frac{1}{2}\right) \xrightarrow[\text{law}]{} N\left(0, \frac{1}{\lambda}\sigma_1^2 + \frac{1}{1-\lambda}\sigma_2^2\right) \tag{5.14}$$

where, if we define $F^0HI = G^0HI \equiv L$,

$$\begin{aligned} \sigma_1^2 &= \int_0^1 I(L^{-1}(z))z^2\, dz - \frac{1}{4}, \\ \sigma_2^2 &= \int_0^1 H(L^{-1}(z))z^2\, dz - \frac{1}{4}. \end{aligned} \tag{5.15}$$

Since $L(s) \leqq H(s)$ and $L(s) \leqq I(s)$ for all values of s, we have

$$\int_0^1 z^3\,dz - \frac{1}{4} \leqq \sigma_1^2, \qquad \sigma_2^2 \leqq \int_0^1 z^2\,dz - \frac{1}{4}, \tag{5.16}$$

or

$$0 \leqq \sigma_1^2, \qquad \sigma_2^2 \leqq \frac{1}{12}. \tag{5.17}$$

The upper bound is attained when there is no censoring, and coincides, of course, with the usual asymptotic variance of the Wilcoxon statistic.

If the censoring distributions H and I are not known to us, it is natural to replace them by their sample estimates, $\hat{H}$ and $\hat{I}$. Substituting into (5.13) yields the statistic

$$W_{\hat{c}} = -\int_{-\infty}^{\infty} \hat{F}(s)\hat{I}(s)\,d[\hat{G}(s)\hat{H}(s)]. \tag{5.18}$$

This statistic can also be derived from permutation considerations: define

$$Q(x_i, y_j, u_k, v_\ell) = \begin{cases} 1 \text{ if } \min(x_i, v_\ell) \geqq \min(y_j, u_k), \\ 0 \text{ if } \min(x_i, v_\ell) < \min(y_j, u_k). \end{cases} \tag{5.19}$$

Then

$$W_{\hat{c}} = \frac{1}{m^2n^2} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{m} \sum_{\ell=1}^{n} Q(x_i, y_j, u_k, v_\ell). \tag{5.20}$$

$W_{\hat{c}}$ is asymptotically normal, by the multisample U statistic theorem, and has expectation $1/2$ under H_0. In some ways, it might be tempting to use (5.18) instead of (5.13) even if H and I were completely known, since given any sample, the randomness of the censoring variables is really of no interest to us. On the other hand, to compute $W_{\hat{c}}$, we must know *all* the u_i and v_j values, as opposed to W_G, W_c, and, as it will turn out, $\hat{W}$, where only the u_i and v_j corresponding to censored x_i and y_j are needed. Even if all the u_i and v_j are available, which is often not the case, this is philosophically unsatisfying, particularly from a Bayesian point of view.

As commented before, the W_c test will be shown to have low efficiency on the case considered in section 9. This is mainly because it ignores the difference between censored and uncensored x_i and y_j, as can be seen in (5.13). It is easy to construct cross censorship tests which do not ignore the differences, but the author has not computed efficiencies for any such test.

6. Alternatives to the null hypothesis

A desirable property of any test statistic for the two sample problem is that when the null hypothesis is not true, that is when $F^0 \neq G^0$, the statistic estimates some reasonable measure of the difference between the two distributions. Usually, we are not interested in a simple acceptance or rejection of the null hypothesis, but would like to make a quantitative assessment of the treatment differences. One of the virtues of the usual Wilcoxon statistic, as applied

to uncensored data, is that it estimates $P\{F^0 \geqq G^0\}$, a parameter that is usually very relevant to the investigator. The asymptotic and approximate nature of the tests we are investigating intensifies the need for a test statistic that is also a reasonable estimator. And also, of course, for an asymptotically normal statistic the nature of the expectation when $F^0 \neq G^0$ determines what type of alternatives to the null hypothesis the tests will tend to have the most power against.

The statistics we have discussed so far, W_G, W_c and $W_{\hat{c}}$ have the unfortunate property that when $F^0 \neq G^0$, their expectations depend on the censoring distributions H and I, which play the role of nuisance parameters in our nonparametric situation. Concentrating our attention on W_G, let us write $x_i^0 \gg y_j^0$ if we can infer from the available data that $x_i^0 \geqq y_j^0$, which will be the case if and only if $x_i \geqq y_j$ and $\epsilon_j = 1$. Likewise, write $y_j^0 \gg x_k^0$ if $y_j > x_i$ and $\delta_i = 1$. Equation (3.4) can then be rewritten as

$$EW_G = \frac{1}{2}[P\{F^0HI \geqq G^0\} - P\{G^0HI > F^0\}] + \frac{1}{2} \tag{6.1}$$

$$= \frac{1}{2}[P\{x_i^0 \gg y_j^0\} - P\{y_j^0 \gg x_i^0\}] + \frac{1}{2}.$$

We see that the W_G test will always be consistent when F^0 is stochastically larger (or smaller) than G^0, for then $F^0(s) > G^0(s)$ for all s, and

$$\begin{aligned} &P\{F^0HI \geqq G^0\} - P\{G^0HI > F^0\} \\ &= -\int_{-\infty}^{\infty} F^0(s)H(s)I(s)\,dG^0(s) + \int_{-\infty}^{\infty} G^0(s)H(s)I(s)\,dF^0(s) \\ &= -\int_{-\infty}^{\infty} F^0HI\,d(G^0 - F^0) + \int_{-\infty}^{\infty} (G^0 - F^0)HI\,dF^0 \\ &= \int_{-\infty}^{\infty} (G^0 - F^0)\,d(F^0HI) + \int_{-\infty}^{\infty} (G^0 - F^0)HI\,dF^0 > 0, \end{aligned} \tag{6.2}$$

which implies from (6.1) that $EW_G > 1/2$.

In cases where F^0 and G^0 are not stochastically comparable, the parameter EW_G can be positive, negative, or zero, depending on the censoring distributions, and even in simple situations can yield misleading information. Consider for example, the case where the x_i^0 and the u_i have independent uniform distributions over the interval $(-1, 1)$, while the y_j^0 and v_j are uniformly distributed over $(-1/2, 1/2)$. A simple calculation shows that $P\{x_i^0 \gg y_j^0\} = 17/96$, while $P\{z_j^0 \gg x_i^0\} = 31/96$; that is, among the expected 50 per cent of the (x_i, y_j) pairs where we know the ordering of (x_i^0, y_j^0), nearly twice as many will favor y_j^0 as favor x_i^0. The expectation of W_G is 0.426 in this case, indicating a strong advantage for the y^0 distribution. Such a conclusion would obviously be inappropriate in many situations.

One of the major advantages of the $\hat{W}$ statistic introduced in section 8 is that it estimates the usual Wilcoxon parameter $P\{F^0 \geqq G^0\}$, independently of the censoring distributions H and I.

7. Nonparametric estimates of F^0 and G^0

An obvious weakness in the definition of W_G can be seen in formula (3.3) in those cases where we are not certain of the ordering between x_i^0 and y_j^0 we ascribe a probability of 1/2 to both possible cases $x_i^0 \geqq y_j^0$, $x_i^0 < y_j^0$, irrespective of the relative magnitudes of x_i and y_j. We will now discuss nonparametric estimates of F^0 and G^0 with an eye toward remedying this weakness. Our discussion will be from a point of view particularly suitable to the two sample problem.

Given $x_1, x_2, \cdots, x_m$ and $\delta_1, \delta_2, \cdots, \delta_m$, we have already introduced $\hat{F}(s)$, the sample estimate for $F(s)$, as the usual right sided empirical c. d. f. of the x_i. That is, letting

$$N_x(s) = \text{number of } x_i \geqq s, \tag{7.1}$$

we define

$$\hat{F}(s) = \frac{N_x(s)}{m}, \tag{7.2}$$

a left continuous, nonincreasing function of s satisfying $\hat{F}(-\infty) = 1$, $\hat{F}(\infty) = 0$, and representing a discrete probability distribution with mass $1/m$ at $x_1, x_2, \cdots, x_m$.

Similarly, we would like to define $\hat{F}^0(s) = N_{x^0}(s)/m$, when $N_{x^0}(s)$ equals the number of $x_i^0 \geqq s$, but because of the censorship, the function $N_{x^0}(s)$ is not available to us. Since $x_i^0 \geqq x_i$ for every i, we do know that $x_i \geqq s$ implies $x_i^0 \geqq s$, so $N_{x^0}(s) \geqq N_x(s)$ for every s. For an $x_i < s$ that is uncensored, $x_i^0 = x_i < s$, and x_i^0 cannot contribute to $N_{x^0}(s)$. The ambiguous situation is $x_i < s$, x_i censored ($\delta_i = 0$), in which case x_i^0 will be equal to or greater than s with conditional probability $F^0(s)/F^0(x_i)$. Of course we do not know $F^0(s)$, but given any initial estimate of it, say $\hat{F}_1^0(s)$, it seems natural to estimate the conditional probability $P\{x_i^0 \geqq s | x_i < s, \delta_i = 0\}$ by $\hat{F}_1^0(s)/\hat{F}_1^0(x_i)$, and define an improved estimate of $F^0(s)$ by

$$\begin{aligned} m\hat{F}_2^0(s) &= N_x(s) + \sum_{\substack{x_i < s \\ \delta_i = 0}} \frac{\hat{F}_1^0(s)}{\hat{F}_1^0(x_i)}, \\ &= N_x(s) + \sum_{x_i < s} (1 - \delta_i) \frac{\hat{F}_1^0(s)}{\hat{F}_1(x_i)}. \end{aligned} \tag{7.3}$$

Iterating, we could then use $\hat{F}_2^0$ in place of $\hat{F}_1^0$ above to get another improved estimate $\hat{F}_3^0$, and so forth, and the question arises whether the sequence $\hat{F}_1^0$, $\hat{F}_2^1$, $\hat{F}_3^0$, $\cdots$ would converge to a function $\hat{F}^0$ which could then not be further improved by application of (7.3). Such a function would have to satisfy

$$m\hat{F}^0(s) = N_x(s) + \sum_{x_i < s} (1 - \delta_i) \frac{\hat{F}^0(s)}{\hat{F}^0(x_i)}, \tag{7.4}$$

for all s. We shall call a function satisfying (7.4) a "self-consistent" estimate of $F^0(s)$, and prove the following theorem.

THEOREM 7.1. *Given* $x_1, x_2, \cdots, x_m$ *and* $\delta_1, \delta_2, \cdots, \delta_m$, *there is a unique function* $\hat{F}^0(x)$ *satisfying* (7.4) *for all values of* s. *Assume without loss of generality, that* $x_1 < x_2 < x_3 < \cdots < x_m$. *Then* $\hat{F}^0(s)$ *has all the usual properties of a right c. d. f., and represents a discrete distribution with mass*

$$\left[m - \sum_{i=1}^{k-1} \frac{1 - \delta_i}{\hat{F}^0(x_i)}\right]^{-1} \tag{7.5}$$

at x_k *if* x_k *is uncensored, and mass*

$$\left[m - \sum_{i=1}^{m-1} \frac{1 - \delta_i}{\hat{F}^0(x_i)}\right]^{-1} \tag{7.6}$$

at x_m *uncensored or not. Here* $\hat{F}^0(s)$ *is defined iteratively by*

$$\hat{F}^0(s) = \begin{cases} 1, & s \leqq x_1, \\ \dfrac{N_x(s)}{\left[m - \sum_{i=1}^{k-1} \dfrac{1 - \delta_i}{\hat{F}^0(x_i)}\right]}, & x_{k-1} < s \leqq x_k, \quad 2 \leqq k \leqq m, \\ 0, & s > x_m. \end{cases} \tag{7.7}$$

COROLLARY 7.1. *The self-consistent estimate* $\hat{F}^0(s)$ *coincides with Kaplan and Meier's* [7] *product limit estimate*

$$\hat{P}(s) = \prod_{i=1}^{k-1} \left(\frac{m - i}{m - i + 1}\right)^{\delta_i}, \qquad s \in (x_{k-1}, x_k], \tag{7.8}$$

if we define $\hat{P}(s) = 0$ *for* $s > x_m$. They have shown [7] that $\hat{P}(s)$ is the nonparametric maximum likelihood estimate of $F^0(s)$.

PROOF. Consider the function $\hat{F}^0(s)$ defined by (7.7). For $s \leqq x_1$, $\hat{F}^0(s) = 1$ and satisfies (7.4). If $\delta_1 = 1$, we see that $\hat{F}^0(x_1) - \hat{F}^0(x_1+) = 1/m$. If $\delta_1 = 0$, $m - (1 - \delta_1)/\hat{F}^0(x_1) = m - 1$, and $\hat{F}^0(x_1+) = \hat{F}^0(x_1) = 1$.

Suppose now that for some k, with $2 \leqq k \leqq m - 1$,

$$m - \sum_{i=1}^{k-1} \frac{1 - \delta_i}{\hat{F}^0(x_i)} > 0, \qquad \hat{F}^0(x_{k-1}+) > 0. \tag{7.9}$$

Then from (7.7)

$$\left[m - \sum_{i=1}^{k-1} \frac{1 - \delta_i}{\hat{F}^0(x_i)}\right] \hat{F}^0(s) = N_x(s), \qquad x_{k-1} < s \leqq x_k, \tag{7.10}$$

and $\hat{F}^0(s)$ satisfies (7.4) for s in the half open interval $(x_{k-1}, x_k]$. Since $N_x(s)$ is constant in this interval, so is $\hat{F}^0(s)$, and $\hat{F}^0(x_k) = \hat{F}^0(x_{k-1}+)$. If $\delta_k = 1$,

$$m - \sum_{i=1}^{k} \frac{1 - \delta_i}{\hat{F}^0(x_i)} = m - \sum_{i=1}^{k-1} \frac{1 - \delta_i}{\hat{F}^0(x_i)} > 0, \tag{7.11}$$

$$\hat{F}^0(x_k+) = [N_x(x_k) - 1]\left[m - \sum_{1}^{k-1} \frac{1 - \delta_i}{\hat{F}^0(x_i)}\right] > 0, \tag{7.12}$$

with

$$\hat{F}^0(x_k) + \hat{F}^0(x_k+) = \left[m - \sum_{i=1}^{k-1} \frac{1 - \delta_i}{\hat{F}^0(x_i)}\right]^{-1}. \tag{7.13}$$

If $\delta_k = 0$,

$$\left[m - \sum_{i=1}^{k} \frac{1-\delta_i}{\hat{F}^0(x_i)}\right]\hat{F}^0(x_i) = \left[m - \sum_{i=1}^{k-1} \frac{1-\delta_i}{\hat{F}^0(x_i)}\right]\hat{F}^0(x_i) - 1 = N_x(x_k) - 1 = N_x(x_k+), \tag{7.14}$$

which implies that (7.11) > 0 in this case. Since by (7.7)

$$\left[m - \sum_{i=1}^{k} \frac{1-\delta_i}{\hat{F}^0(x_i)}\right]\hat{F}^0(x_k+) = N_x(x_k+), \tag{7.15}$$

we have $\hat{F}^0(x_k+) = \hat{F}^0(x_k) > 0$.

In either case, the argument proceeds by induction until we have verified (7.4) for $s \in (-\infty, x_{m-1}]$, and shown that $\hat{F}^0(x_{m-1}+) > 0$, $m - \sum_{i=1}^{m-1}(1-\delta_i)/\hat{F}^0(x_i) = 0$. From (7.7), this verifies (7.4) for $s \in (x_{m-1}, x_m]$, while $F^0(s) = 0$ trivially satisfies (7.4) for $s > x_m$. Since $\hat{F}^0(x_m) = \hat{F}^0(x_{m-1}+)$, the jump at x_m equals (7.5). Note that if $\delta_m = 1$,

$$m - \sum_{i=1}^{m} \frac{1-\delta_i}{\hat{F}^0(x_i)} = m - \sum_{i=1}^{m-1} \frac{1-\delta_i}{\hat{F}^0(x_i)} > 0, \tag{7.16}$$

while if $\delta_m = 0$,

$$\left[m - \sum_{1}^{m} \frac{1-\delta_i}{\hat{F}^0(x_i)}\right]F_0(x_i) = N_x(x_m) - 1 = 0, \tag{7.17}$$

implying $m - \sum_1^m (1-\delta_i)/\hat{F}^0(x_i) = 0$.

The argument verifying the uniqueness of $\hat{F}^0(s)$ proceeds by induction in a manner almost identical with the above.

To prove the corollary, note that $\hat{F}^0(s) = \hat{P}(s) = 1$, for $s \leqq x_1$. Suppose that $\hat{F}^0(x_k) = \hat{P}(x_k)$ for some value of k, with $1 \leqq k \leqq m-1$. If $\delta_k = 0$, then by the proof above and by definitions (7.7) and (7.8) $\hat{F}^0(x_k+) = \hat{F}^0(x_k) = \hat{P}(x_k) = \hat{P}(x_k+)$, implying that $\hat{F}^0(s) = \hat{P}(s)$ for $s \in (x_k, x_{k+1}]$.

If $\delta_k = 1$, then

$$\hat{F}(x_k+) = \frac{N_x(x_k+)}{m - \sum_1^k \frac{1-\delta_i}{\hat{F}^0(x_i)}} = \frac{N_x(x_k+)}{N_x(x_k)} \frac{N_x(x_k)}{m - \sum_1^{k-1} \frac{1-\delta_i}{\hat{F}^0(x_i)}} = \frac{m-k}{m-k+1}\hat{P}(x_k) = \hat{P}(x_k+), \tag{7.18}$$

implying again that $\hat{F}^0(s) = \hat{P}(s)$ for $s \in (x_k, x_{k+1}]$. The proof of the corollary is completed by induction.

The self-consistent (product limit) estimate is also given by the following simple construction; place probability mass $1/m$ at each of the points $x_1 < x_2 < x_3 < \cdots < x_m$, (that is, construct the distribution with $\hat{F}(s)$ as c. d. f.); if x_{i_1} is the smallest x_i that is censored, remove the mass at x_{i_1} and redistribute it equally among the $m - i_1$ points to the right of it, $x_{i_1+1}, x_{i_1+2}, \cdots, x_m$. If x_{i_2} is the smallest censored value among $x_{i_1+1}, x_{i_1+2}, \cdots, x_m$, redistribute its mass,

which will be $1/m + 1/(m - i_1)$, among the $m - m_2$ points to its right, x_{i_2+1}, $x_{i_2+2}, \cdots, x_m$; continue in this way until you reach x_m. The resulting distribution has mass

$$\frac{1}{m}\prod_{1}^{k-1}\left(\frac{m-i+1}{m-i}\right)^{1-\delta_i} \tag{7.19}$$

at x_k uncensored, and at x_m censored or not, which is easily computed to agree with $\hat{F}^0(s)$ from equation (7.8). This construction sheds some light on the special nature of the largest observation, which the self-consistent estimator always treats as uncensored, irrespective of δ_m.

In addition to the maximum likelihood property of $\hat{F}^0$, Kaplan and Meier derive several other results which will be of use to us in the next section.

(a) $\hat{F}^0(s)$ is a nearly unbiased estimator of $F^0(s)$. Specifically,

$$1 \geqq \frac{E\hat{F}^0(s)}{F^0(s)} \geqq 1 - e^{-EN_x(s)} \tag{7.20}$$

or equivalently,

$$0 \leqq F^0(s) - E\hat{F}^0(s) \leqq F^0(s)e^{-EN_x(s)}. \tag{7.21}$$

Since $EN_x(s) = mF(s)$, the bias of $\hat{F}^0(s)$ declines exponentially with sample size whenever $F(s) = F^0(s)H(s) > 0$ (that is, whenever it is possible and necessary to estimate $F^0(s)$ nonparametrically). In the sequel, we will treat $\hat{F}^0(s)$ as if it were unbiased, since the exponential decline of the bias term easily overwhelms the $m^{1/2}$ magnification needed for the large sample theory.

Harking back, briefly, to the point raised in section 2, equation (7.20) also holds when the u_i are a fixed set of constants, say $u_1 < u_2 < u_3 < \cdots < u_m$, in which case it is simple to show that the relative bias $E\hat{F}^0(s)/F^0(s)$ is a constant within any interval $(u_{k-1}, u_k]$.

(b) $\hat{F}^0(s)$ is a consistent estimator of $F^0(s)$, as m goes to infinity.

(c) $m^{1/2}[\hat{F}^0(s) - F^0(s)]$, considered as a stochastic process in s, approaches in the limit for large m, a normal process with mean 0 and covariance kernel

$$\Gamma(s, t) = F^0(s)F^0(t)\int_{-\infty}^{s}\frac{-dF^0(z)}{F(z)F^0(z)}, \qquad s \leqq t. \tag{7.22}$$

(The limiting normality is not discussed by Kaplan and Meier, but can be derived easily from (7.7) by standard methods.)

The covariance kernel (7.9) represents a distorted Wiener process of the type discussed by Doob in his famous paper on the Kolmogorov-Smirnov statistic [8]. Suppose, for convenience, that $F^0(s) = 1$, so that we are dealing with positive random variables. Define

$$a(s) = \int_0^s \frac{-dF^0(z)}{F(z)F^0(z)}, \tag{7.23}$$

an increasing function of s for s less than M, any value such that $F^0(M) > 0$, and let $a^{-1}(s)$ denote its inverse function. Then for $0 \leqq s \leqq M$, the process

$$Y(s) = m^{1/2} \frac{\hat{F}^0[a^{-1}(s)] - F^0[a^{-1}(s)]}{F^0[a^{-1}(s)]} \tag{7.24}$$

approaches a standard Wiener process as m goes to infinity. That is, Y approaches a normal process with mean zero and covariance kernel $\Gamma(s, t) = \min(s, t)$ for $0 \leqq s, t \leqq M$.

8. The statistic $\hat{W}$

We now use $\hat{F}^0(s)$ and $\hat{G}^0(s)$, the self-consistent estimates of F^0 and $\hat{G}^0$, respectively, to modify the statistic W_G along the lines suggested at the beginning of section 7. As mentioned previously, the self-consistent estimates treat the largest observation of each group as if it were uncensored, and we will assume that this is actually the situation, to avoid a host of annoying special cases. That is, we will assume the δ_i and e_j corresponding to the largest x_i and largest y_j respectively are both equal to 1.

Define the scoring function $\hat{Q}(x_i, y_j, \delta_i, \epsilon_j)$ to be

$$\hat{Q}(x_i, y_j, \delta_i, \epsilon_j) = P\{x_i^0 \geqq y_j^0 | x_i, y_j, \delta_i, \epsilon_j, \hat{F}^0, \hat{G}^0\}, \tag{8.1}$$

where the conditional expectation is interpreted as if x_i^0 and y_j^0 were actually drawn from $\hat{F}^0$ and $\hat{G}^0$, respectively. Thus, if $x_i \geqq y_j$ and $\epsilon_j = 1$, then $\hat{Q}(x_i, y_j, \delta_i, \epsilon_j) = 1$ as before. However, if $x_i \geqq y_j$ and $\epsilon_j = 0$, $\delta_i = 1$, we no longer score 1/2 to indicate equal probability for $x_i^0 \geqq y_j^0$ and $x_i^0 < y_j^0$; rather $\hat{Q}(x_i, y_j, \delta_i, \epsilon_j) = 1 - \hat{G}^0(x_i)/\hat{G}^0(y_j)$ in this case, the conditional probability under $\hat{G}^0$ that $y_j^0 \geqq y_j$ is less than $x_i^0 = x_i$. Table I lists the value of $Q(x_i, y_j, \delta_i, \epsilon_j)$ in all eight possible different cases.

TABLE I

VALUES OF $Q(x_i, y_j, \delta_i, \epsilon_j)$

(δ_i, ϵ_j)	$x_i \geqq y_j$	$x_i < y_j$
(1, 1)	1	0
(0, 1)	1	$\frac{\hat{F}^0(y_j)}{\hat{F}^0(x_i)}$
(1, 0)	$1 - \frac{\hat{G}^0(x_i)}{\hat{G}^0(y_j)}$	0
(0, 0)	$1 - \frac{\hat{G}^0(x_i)}{\hat{G}^0(y_j)} - \int_{x_i}^{\infty} \frac{F^0(s)\, dG^0(s)}{F^0(x_i)G^0(y_j)}$	$-\int_{y_i}^{\infty} \frac{\hat{F}^0(s)\, \alpha\hat{G}^0(s)}{\hat{F}^0(x_i)\hat{G}^0(y_j)}$

The statistic $\hat{W}$ is now defined in the usual way in terms of $\hat{Q}(x_i, y_j, \delta_i, \epsilon_j)$

$$\hat{W} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \hat{Q}(x_i, y_j, \delta_i, \epsilon_j). \tag{8.2}$$

THEOREM 8.1. $\hat{W} = -\int_{-\infty}^{\infty} \hat{F}^0(s)\, d\hat{G}^0(s) = P\{\hat{F}^0 \geqq \hat{G}^0\}$, *and is the maximum likelihood estimate of* $P\{\hat{F}^0 \geqq \hat{G}^0\}$.

PROOF. To simplify notation, let $P\{x_i^0 \geqq y_j^0\}$ be the conditional expectation defined at (8.1) and below, and listed in table I. We shall also need $\hat{g}^0(y_j) = \hat{G}^0(y_j) - \hat{G}^0(y_j+)$, the probability mass function corresponding to the discrete distribution $\hat{G}^0(s)$. From theorem 7.1

$$\hat{g}^0(y_j) = \frac{\epsilon_j}{n - \sum_{y_\ell < y_i} \dfrac{1 - \epsilon_\ell}{\hat{G}^0(y_\ell)}}, \tag{8.3}$$

for $j = 1, 2, 3, \cdots, n$, which can be written as

$$n\hat{g}^0(y_j) = 1 + \sum_{y_\ell < y_i} \frac{1 - \epsilon_\ell}{\hat{G}^0(y_\ell)} \hat{g}^0(y_j), \qquad \epsilon_j = 1. \tag{8.4}$$

If $\epsilon_j = 1$, then reading from the table,

$$\begin{aligned} \sum_{i=1}^{m} \hat{P}\{x_i^0 \geqq y_j^0\} &= N_x(y_j) + \sum_{x_i < y_i} (i - \delta_i) \frac{\hat{F}^0(y_j)}{\hat{F}^0(x_i)} \\ &= m\hat{F}^0(y_j) \end{aligned} \tag{8.5}$$

by the self-consistency property (7.4) of $\hat{F}^0$.

If $\epsilon_j = 0$, then

$$\hat{P}\{x_i^0 \geqq y_j^0\} = \sum_{y_\ell \geqq y_i} \frac{\hat{g}^0(y_\ell)}{\hat{G}^0(y_j)} \hat{P}\{x_i^0 \geqq y_\ell^0\}, \tag{8.6}$$

so

$$\begin{aligned} \sum_{i=1}^{m} \hat{P}\{x_i^0 \geqq y_j^0\} &= \sum_{i=1}^{m} \sum_{y_\ell \geqq y_i} \frac{\hat{g}^0(y_\ell)}{\hat{G}^0(y_j)} \hat{P}\{x_i^0 \geqq y_\ell^0\} \\ &= \sum_{y_\ell \geqq y_i} \frac{\hat{g}^0(y_\ell)}{\hat{G}^0(y_j)} \left[\sum_{i=1}^{m} \hat{P}\{x_i^0 \geqq y_\ell^0\} \right]. \end{aligned} \tag{8.7}$$

Remembering that $\hat{g}^0(y_\ell)$ is nonzero only for uncensored values of y_ℓ, equation (8.5) reduces this last sum to

$$\sum_{i=1}^{m} \hat{P}\{x_i^0 \geqq y_j^0\} = m \sum_{y_\ell \geqq y_i} \frac{\hat{g}^0(y_\ell)}{\hat{G}^0(y_j)} \hat{F}^0(y_\ell). \tag{8.8}$$

Combining (8.5) and (8.8) yields

$$\begin{aligned} \hat{W} &= \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \hat{P}\{x_i^0 \geqq y_j^0\} \\ &= \frac{1}{n} \left[\sum_{\epsilon_j = 1} \hat{F}^0(y_j) + \sum_{\epsilon_j = 0} \sum_{y_\ell \geqq y_i} \frac{\hat{g}^0(y_\ell)}{\hat{G}^0(y_j)} \hat{F}^0(y_\ell) \right] \\ &= \frac{1}{n} \sum_{\epsilon_j = 1} \hat{F}^0(y_j) \left[1 + \sum_{y_\ell < y_i} \frac{1 - \epsilon_\ell}{\hat{G}^0(y_\ell)} \hat{g}^0(y_j) \right] \\ &= \sum_{\epsilon_j = 1} \hat{F}^0(y_j) \hat{g}^0(y_j) \end{aligned} \tag{8.9}$$

by equation (8.4). Thus,

$$\hat{W} = -\int_{-\infty}^{\infty} \hat{F}^0(s)\, d\hat{G}^0(s) = P\{\hat{F}^0 \geqq \hat{G}^0\}, \tag{8.10}$$

as was to be proved, and the second statement of the theorem follows from the corollary to theorem 7.1, and the usual invariance property of maximum likelihood estimates.

It should be mentioned that other consistent estimates of F^0 and G^0 exist [7], which conceivably could be used in place of $\hat{F}$ and $\hat{G}$ to define the scoring function, as was done in (8.1). However, the uniqueness of the self-consistent estimate insures that the resulting statistic will not be expressible in the form given in theorem 8.1.

THEOREM 8.2. *Let m and n go to infinity in such a way that* $\lim m/(m+n) = \lambda$, *with* $0 < \lambda < 1$. *Then*

$$(m+n)^{1/2}[\hat{W} - P\{F^0 \geqq G^0\}] \underset{\text{law}}{\to} N\left(0, \frac{1}{\lambda}\sigma_1^2 + \frac{1}{1-\lambda}\sigma_2^2\right) \tag{8.11}$$

where under H_0,

$$\begin{aligned} \sigma_1^2 &= \frac{1}{4}\int_0^1 \frac{z^2\, dz}{H[F^{0^{-1}}(z)]} = \frac{1}{4}\int_0^1 \frac{z^3\, dz}{F[F^{0^{-1}}(z)]}, \\ \sigma_2^2 &= \frac{1}{4}\int_0^1 \frac{z^2\, dz}{I[G^{0^{-1}}(z)]} = \frac{1}{4}\int_0^1 \frac{z^3\, dz}{G[G^{0^{-1}}(z)]}. \end{aligned} \tag{8.12}$$

Here $F^{0^{-1}}$ and $G^{0^{-1}}$ are the inverse functions of F^0 and G^0, respectively, and are identical under H_0.

PROOF. Only a heuristic argument will be presented here. This argument is somewhat similar to the proof of the Chernoff-Savage theorem [9], and a rigorous proof can be developed along the lines suggested in that paper.

Write

$$\begin{aligned} -\hat{W} &= \int_{-\infty}^{\infty} \hat{F}^0(s)\, d\hat{G}^0(s) \\ &= \int_{-\infty}^{\infty} F^0\, dG^0 + \int_{-\infty}^{\infty} (\hat{F}^0 - F^0)\, dG^0 + \int_{-\infty}^{\infty} F^0\, d(\hat{G}^0 - \hat{G}) \\ &\quad + \int_{-\infty}^{\infty} (\hat{F}^0 - F^0)\, d(\hat{G}^0 - \hat{G}), \end{aligned} \tag{8.13}$$

and integrate the third term by parts, yielding

$$\begin{aligned} -\hat{W} &= \int_{-\infty}^{\infty} F^0\, dG^0 + \int_{-\infty}^{\infty} (\hat{F}^0 - F^0)\, dG^0 - \int_{-\infty}^{\infty} (\hat{G}^0 - G^0)\, dF^0 \\ &\quad + \int_{-\infty}^{\infty} (\hat{F}^0 - F^0)\, d(\hat{G}^0 - G^0). \end{aligned} \tag{8.14}$$

Of these four terms, the first is a constant, while the fourth is asymptotically negligible;

$$(8.15)\qquad \int_{-\infty}^{\infty} (\hat{F}^0 - F^0)\, d(\hat{G}^0 - G^0)$$
$$= \frac{1}{(mn)^{1/2}} \int_{-\infty}^{\infty} m^{1/2}(\hat{F}^0 - F^0)\, dn^{1/2}(\hat{G}^0 - G^0)$$
$$= o_P\left[\frac{1}{(m+n)^{1/2}}\right]$$

by property c of section 7. (The equation $x_{m,n} = o_P[1/(m+n)^{1/2}]$ means $(m+n)^{1/2}x_{m,n} \xrightarrow[\text{prob}]{} 0$ as m, n go to infinity.)

Consider the second term in (8.14). By property c of section 7 and equation (7.18), $\int_{-\infty}^{\infty} m^{1/2}[\hat{F}^0(s) - F^0(s)]\, dG^0(s)$ approaches a normal variate with mean zero and variance

$$(8.16)\qquad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Gamma(s,t)\, dG^0(s)\, dG^0(t)$$
$$= -2\int_{-\infty}^{\infty}\int_{s}^{\infty}\int_{-\infty}^{s} \frac{F^0(s)F^0(t)}{F(z)F^0(z)}\, dF^0(z)\, dG^0(t)\, dG^0(s).$$

Now under the null hypothesis, $G^0 = F^0$, this last expression becomes

$$(8.17)\qquad -2\int_{-\infty}^{\infty}\int_{s}^{\infty}\int_{-\infty}^{s} \frac{F^0(s)F^0(t)}{F(z)F^0(z)}\, dF^0(z)\, dF^0(t)\, dF^0(s),$$

and a change of variables to $F^0(z)$, $F^0(t)$ and $F^0(s)$ yields

$$(8.18)\qquad 2\int_0^1\int_0^s\int_s^1 \frac{st}{zF[F^{0-1}(z)]}\, dz\, dt\, ds$$
$$= 2\int_0^1\int_0^z\int_0^s \frac{st}{zF[F^{0-1}(z)]}\, dt\, ds\, dz$$
$$= \frac{1}{4}\int_0^1 \frac{z^3\, dz}{F[F^{0-1}(z)]} = \frac{1}{4}\int_0^1 \frac{z^2\, dz}{H[F^{0-1}(z)]} = \sigma_1^2.$$

The last equality follows from (2.5)

$$(8.19)\qquad F[F^{0-1}(z)] = H[F^{0-1}(z)]F^0[F^{0-1}(z)] = zH[F^{0-1}(z)].$$

A similar argument gives

$$(8.20)\qquad \sigma_2^2 = \frac{1}{4}\int_0^1 \frac{z^2\, dz}{I[G^{0-1}(z)]}$$

as the limiting variance of the third term in (8.13) under H_0. Since the second and third terms of that expression are independent random variables, this completes the heuristic argument.

The expression (8.11) for σ_1^2 fails to converge if $H(z) = O\{[F^0(z)]^3\}$, as z approaches 0, and likewise σ_2^2 will not converge if $I(z) = O\{[G^0(z)]^3\}$. In these cases, the variance of $\hat{W}$ does not diminish at rate $1/(m+n)$. This situation is illustrated in the next section, and a possible remedy suggested in section 10.

In these cases where σ_1^2 and σ_2^2 are finite, they have obvious consistent estimators in terms of equation (8.11). The second form given for each integral will often be the most convenient to compute with, since $\hat{F}^0(z) \geqq \hat{F}(z)$ for all z, and the function $\hat{F}(z)$ can be computed directly from the observed x_i. (To compute $\hat{H}$, for instance, requires knowing all the u_i values, not just those where $x_i = u_i$. As mentioned before, these may not be available, and in any case are not required to compute $\hat{W}$.)

9. Asymptotic relation efficiencies of the various tests in the exponential case

In this section we will compute the Pitman efficiency (A. R. E.) of $\hat{W}$, W_G, and W_c relative to the best parametric test, in the special case where all the random variables involved are exponentially distributed. That is, we assume

$$(9.1)\qquad \begin{aligned} F^0(s) &= \begin{cases} e^{-\phi s}, & s \geqq 0, \\ 1, & s < 0, \end{cases} \qquad G^0(s) = \begin{cases} e^{-\theta\phi s}, & s \geqq 0, \\ 1, & s < 0, \end{cases} \\ H(s) &= \begin{cases} e^{-s}, & s \geqq 0, \\ 1, & s < 0, \end{cases} \qquad I(s) = \begin{cases} e^{-\alpha s}, & s \geqq 0, \\ 1, & s < 0, \end{cases} \end{aligned}$$

where, for the purposes of the parametric test, α is a known positive parameter, while θ and ϕ are unknown. This example with $\alpha = 1$ was investigated by Gilbert. Gehan evaluates the more realistic case where $H(s) = I(s) =$ a uniform distribution over $[0, T]$. Our null hypothesis is $H_0\colon \theta = 1$. The observed random variables x_i and y_j will have right c. d. f.'s,

$$(9.2)\qquad F(s) = \begin{cases} e^{-(\phi+1)s}, & s \geqq 0, \\ 1, & s < 0, \end{cases} \qquad G(s) = \begin{cases} e^{-(\theta\phi+\alpha)s}, & s \geqq 0, \\ 1, & s < 0, \end{cases}$$

respectively. The δ_i and ϵ_j are Bernoulli random variables, with

$$(9.3)\qquad P\{\delta_i = 1\} = \frac{\phi}{\phi + 1}, \qquad P\{\epsilon_j = 1\} = \frac{\theta\phi}{\theta\phi + \alpha}.$$

In this case, x_i and δ_i are independent, as are y_j and ϵ_j. (I am grateful to Dr. J. Sethuraman, for first bringing this useful fact of the independence of x_i and δ_i to my attention.) To simplify matters further, we will let $m = n$, so the sample sizes remain equal as n goes to infinity.

The efficacy $\mathrm{Eff}(T_n)$ of any test statistic T_n for H_0 is defined to be

$$(9.4)\qquad \mathrm{Eff}(T) = \frac{\left(\dfrac{\partial ET_n}{\partial\theta}\Big|_{\theta=1}\right)^2}{2n \operatorname{Var}_{\theta=1,\phi} T}.$$

From the discussion in section 4, we know that the maximum likelihood estimate of θ, say M, achieves the greatest efficacy among all computable test statistics, the class we called C. That efficacy is $1/I^{00}$. In our case, assuming θ, φ unknown, but α known,

$$\text{Eff}(M) = \frac{1}{4}\frac{\phi}{\phi + \frac{\alpha+1}{2}} \tag{9.5}$$

at $\theta = 1$.

From (4.2) and (8.12), the efficacies of W_G and $\hat{W}$ at $\theta = 1$ are

$$\text{Eff}(W_G) = \frac{3}{16}\frac{\phi\left(\phi + \frac{2\alpha+1}{3}\right)\left(\phi + \frac{\alpha+2}{3}\right)}{\left(\phi + \frac{\alpha+1}{2}\right)^3}, \tag{9.6}$$

and

$$\text{Eff}(\hat{W}) = \frac{3}{16}\frac{\left(\phi - \frac{1}{3}\right)\left(\phi - \frac{\alpha}{3}\right)}{\phi\left(\phi - \frac{\alpha+1}{6}\right)}. \tag{9.7}$$

Formula (9.7) holds only for $\phi > \max(1/3, \alpha/3)$. For $\phi \leqq \max(1/3, \alpha/3)$, the situation described at the end of section 8 prevails, and $\hat{W}$ has efficacy 0.

The Pitman efficiency of W_G relative to M, the best parametric test, is by definition

$$\begin{aligned} P_{W_G/M} &= \frac{\text{Eff}(W_G)}{\text{Eff}(M)} \\ &= \frac{3}{4}\frac{\left(\phi + \frac{2\alpha+1}{3}\right)\left(\phi + \frac{\alpha+2}{3}\right)}{\left(\phi + \frac{\alpha+1}{2}\right)^2}. \end{aligned} \tag{9.8}$$

When $\alpha = 1$, that is, when $H(s) = I(s) = e^{-s}$ for $s \geqq 0$, then $P_{W_G/M} = 3/4$ for all values of ϕ, agreeing with Gilbert's result. For all other values of α and ϕ,

$$\frac{8}{9}\left(\frac{3}{4}\right) < P_{W_G/M} < \frac{3}{4} \tag{9.9}$$

with the lower bound achieved at the extreme points $\phi = 0$, $\alpha = 0$ and $\phi = 0$, $\alpha = \infty$.

For the $\hat{W}$ statistic we have Pitman efficiency

$$\begin{aligned} P_{\hat{W}/M} &= \frac{\text{Eff}(\hat{W})}{\text{Eff}(M)} \\ &= \begin{cases} \frac{3}{4}\dfrac{\left(\phi - \frac{1}{3}\right)\left(\phi - \frac{\alpha}{3}\right)\left(\phi + \frac{\alpha+1}{2}\right)}{\phi^2\left(\phi - \frac{\alpha+1}{6}\right)} & \text{for } \phi > \max(1/3, \alpha/3) \\ 0 & \text{for } \phi \leqq \max(1/3, \alpha/3). \end{cases} \end{aligned} \tag{9.10}$$

Let us consider, the convenience, the case where $\alpha \leqq 1$. Then for $\phi \geqq 1/2$, we have $P_{\hat{W}/M} \geqq 3/4$. Another way of saying this is that for exponential observations and exponential censoring variables, the W test is more efficient than the W_G test as long as not more than 2/3 of either sample consists of

censored observations (for when $\phi = 1/2$, an expected $1/(\phi + 1) = 2/3$ percentage of the x_i and $\alpha/(\phi + \alpha) \leqq 2/3$ percentage of the y_j will be censored under H_0: $\theta = 1$). That the increase in efficiency can be quite substantial is shown in table II.

TABLE II

$P_{W_G/M}$ IN THE EXPONENTIAL CASE

ϕ \ α	1	0.5	0.25	0
∞	0.750	0.750	0.750	0.750
10	0.798	0.786	0.780	0.774
5	0.840	0.819	0.808	0.797
3	0.889	0.858	0.841	0.823
2	0.938	0.900	0.877	0.852
1	1.000	0.972	0.941	0.900
7/8	0.995	0.977	0.945	0.901
3/4	0.972	0.972	0.940	0.893
2/3	0.937	0.956	0.925	0.875
1/2	0.750	0.833	0.803	0.750
1/3	0.000	0.000	0.000	0.000

The case of very large ϕ corresponds to a very low rate of censoring $1/(1 + \phi)$, in which case both W_G and W become equivalent to the ordinary Wilcoxon test, while the F test based on $\sum_1^n x_i / \sum_1^n y_j$ corresponds to M. This accounts for the constant top row of the table. The same remark applies to the statistic W_e discussed in section 5, which by (5.14) has Pitman efficiency

$$P_{W_e/M} = \frac{3}{4} \frac{\phi\left(\phi + \frac{5}{3}\right)}{(\phi + 2)^2}, \tag{9.11}$$

when $\alpha = 1$. This is quite low for reasonable values of ϕ, not attaining even 1/2 until ϕ is greater than 4.7.

The last column of the table is for $\alpha = 0$, that is, the case where the y_j^0 values are uncensored.

10. Truncating the $\hat{W}$ statistic

The rapid loss of asymptotic efficiency of the $\hat{W}$ test when ϕ drops below 1/2 in the example treated in section 9 can be alleviated somewhat by considering truncated modifications of the $\hat{W}$ statistic. We will consider here, briefly, such modifications, first in general, and then as applied to the exponential case as given by (9.1).

Let $D(s)$ be any continuous right c. d. f., and define

$$\begin{aligned} x_i^0 &= \min(x_i^0, z_i), \\ y_j^0 &= \min(y_j^0, z_{m+j}). \end{aligned} \tag{10.1}$$

Here the z are mutually independent random variables, identically distributed according to $D(s)$. Following our previous notation, (2.2), (2.3), let

$$\begin{aligned} \tilde{x}_i &= \min(x_i^0, u_i), \qquad & \tilde{y}_j &= \min(y_j^0, v_j), \\ \delta_i &= \begin{cases} 1 \text{ if } \tilde{x}_i = x_i^0, \\ 0 \text{ if } \tilde{x}_i < x_i^0, \end{cases} & \epsilon_j &= \begin{cases} 1 \text{ if } \tilde{y}_j = y_j^0, \\ 0 \text{ if } \tilde{y}_j < y_j^0. \end{cases} \end{aligned} \tag{10.2}$$

These quantities can all be calculated from the data available to the statistician: namely x_i, δ_i and y_j, ϵ_j for $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, n$, and the values of z_i taken from a random number table. For instance, $\delta_i = 1$ implies $\tilde{x}_i = \min(x_i, z_i)$ and $\tilde{\delta}_i = 1$, while $\delta_i = 0$ implies $\tilde{x}_i = \min(x_i, z_i)$ and $\tilde{\delta}_i$ equals 1 or 0 as z_i is less than or greater than x_i. The same considerations apply to $\tilde{y}_j$ and $\tilde{\epsilon}_j$.

The statistician is free to choose any convenient truncation distribution $D(s)$, and apply the $\hat{W}$ test to the modified data (10.2) instead of the original data (2.2), (2.3). The null hypothesis now becomes the equality of the distributions of $\tilde{x}_i^0$ and $\tilde{y}_j^0$, which are $\tilde{F}^0 = DF^0$ and $\tilde{G}^0 = DG^0$, respectively. Applied to the tilde data, the $\hat{W}$ statistic is the maximum likelihood estimate of $P\{DF^0 \geqq DG^0\}$, having that quantity as asymptotic expectation, and variance

$$\frac{1}{m+n}\left(\frac{1}{\lambda}\sigma_1^2 + \frac{1}{1-\lambda}\sigma_2^2\right), \tag{10.3}$$

where $\lambda = \lim m/(m+n)$, and

$$\begin{aligned} \sigma_1^2 &= \frac{1}{4}\int_0^1 \frac{z^2\,dz}{H[(DF^0)^{-1}(z)]}, \\ \sigma_2^2 &= \frac{1}{4}\int_0^1 \frac{z^2\,dz}{I[(DF^0)^{-1}(z)]} \end{aligned} \tag{10.4}$$

under H_0.

Suppose now we choose a $D(s)$ which is very nearly equal to

$$D_T(s) = \begin{cases} 1 \text{ if } s \leqq T, \\ 0 \text{ if } s > T, \end{cases} \tag{10.5}$$

and denote the corresponding statistic by $\hat{W}_T$. Then we have as approximations,

$$E\,\hat{W}_T = \int_{-\infty}^{T} F^0(s)\,d\,G^0(s) + \frac{1}{2}F^0(T)G^0(T), \tag{10.6}$$

and, under H_0,

$$\begin{aligned} \sigma_1^2 &= \frac{[F^0(T)]^3}{12H(T)} + \frac{1}{4}\int_{F^0(T)}^1 \frac{z^2\,dz}{H[F^{0-1}(z)]\,dz}, \\ \sigma_2^2 &= \frac{[F^0(T)]^3}{12I(T)} + \frac{1}{4}\int_{F(T)}^1 \frac{z^2\,dz}{I[F^{0-1}(z)]\,dz}. \end{aligned} \tag{10.7}$$

By choosing T sufficiently small, we can now avoid infinite values of σ_1^2 and σ_2^2.

Applying these results to the exponential case of section 9, with $\alpha = 1$, that is $F^0(s) = e^{-\phi s}$, $G^0(s) = e^{-\theta\phi s}$, $H(s) = I(s) = e^{-s}$, yields

$$E\hat{W}_T = \frac{\theta}{1+\theta} + \left(\frac{1}{2} - \frac{\theta}{1+\theta}\right) e^{-\phi(\theta+1)T} \tag{10.8}$$

$$= \frac{\theta}{1+\theta} + \left(\frac{1}{2} - \frac{\theta}{1+\theta}\right) t^{-(\theta+1)},$$

where we have defined $T = \phi^{-1} \log t$. Under H_0: $\theta = 1$,

$$\sigma_1^2 = \sigma_2^2 = \frac{1}{12} \frac{\phi}{\phi - \frac{1}{3}} \left[1 - \frac{1}{3\phi} t^{-(3-1/\phi)}\right]. \tag{10.9}$$

(Here we are excluding the case $\phi = 1/3$.)

The efficacy of $\hat{W}_T$ is calculated to be

$$\text{Eff}(\hat{W}_T) = \frac{3}{16} \frac{\phi - \frac{1}{3}}{\phi} r(t), \tag{10.10}$$

where

$$r(t) = \frac{(1 - t^{-2})^2}{1 - \frac{1}{3\phi} t^{-(3-1/\phi)}}. \tag{10.11}$$

The Pitman efficiency of the $\hat{W}_T$ test relative to the best parametric test is then

$$P_{\hat{W}_T/M} = \frac{3}{4} \frac{\left(\phi - \frac{1}{3}\right)(\phi + 1)}{\phi^2} r(t). \tag{10.12}$$

Comparing this with

$$P_{\hat{W}/M} = \frac{3}{4} \frac{\left(\phi - \frac{1}{3}\right)(\phi + 1)}{\phi^2} \tag{10.13}$$

for $\phi > 1/3$ (calculated from (9.8) with $\alpha = 1$), shows that truncation can only lower the efficiency when $\phi \geqq 1$, for in this case $r(t) \leqq 1$ for all t. For $\phi < 1$, an improvement in efficiency is possible by truncation at the correct point. The calculations for $\phi = 1/2$ and $\phi = 1/4$ are summarized in table III.

TABLE III

OPTIMUM TRUNCATION POINT AND RELATIVE PITMAN EFFICIENCY CALCULATED FOR $\phi = 1/2$ AND $1/4$

	$\phi = 1/2$	$\phi = 1/4$
Optimum truncation point T	3.38	2.43
Relative truncation point T	1.69	0.61
$P_{\hat{W}_T/M}$	0.798	0.427
$P_{\hat{W}/M}$	0.750	0.000

It should be noted that for $\phi = 1/4$, the greatest power is achieved at the expense of a really drastic truncation, at 61 per cent of the expectation $1/\phi$. A more reasonable procedure, which has not been investigated by the author, is to combine the $\hat{W}$ and W_G tests in those cases where censorship is so severe, say greater than 2/3 of the observations, that the $\hat{W}$ test alone may be unstable. This could be done by using the $\hat{Q}(x_i, y_j, \delta_i, \epsilon_j)$ scoring function in table II for those cases where either x_i or y_j is smaller than a threshold T, and the scoring function Q_G, given by (3.3), in all other cases.

I am very grateful to Lincoln Moses, who has been both generous and accurate with his advice during the course of this work.

REFERENCES

[1] M. HALPERIN, "Extension of the Wilcoxon-Mann-Whitney test to samples censored at the same fixed point," *J. Amer. Statist. Assoc.*, Vol. 55 (1960), pp. 125–138.
[2] E. A. GEHAN, "A generalized Wilcoxon test for comparing arbitrarily singly censored samples," *Biometrika*, Vol. 52 (1965), pp. 203–223.
[3] J. P. GILBERT, "Random censorship," unpublished Ph.D. thesis, University of Chicago, 1962.
[4] N. MANTEL, "Ranking procedures for arbitrarily restricted observation," to appear.
[5] D. A. S. FRASER, *Nonparametric Statistics*, New York, Wiley, 1957, p. 229.
[6] J. NEYMAN, "Optimal asymptotic tests of composite hypothesis," *Probability and Statistics; the Harald Cramér Volume*, New York, Wiley, 1959, pp. 213–235.
[7] E. L. KAPLAN and P. MEIER, "Non-parametric estimation from incomplete observations," *J. Amer. Statist. Assoc.*, Vol. 53 (1958), pp. 457–481.
[8] J. L. DOOB, "Heuristic approach to the Kolmogorov-Smirnov theorems," *Ann. Math. Statist.*, Vol. 120 (1949), pp. 393–402.
[9] H. CHERNOFF and I. R. SAVAGE, "Asymptotic normality and efficiency of certain nonparametric test statistics," *Ann. Math. Statist.*, Vol. 29 (1958), pp. 972–994.