54953

2984

WATER RESOURCES RESEARCH, VOL. 24, NO. 12, PAGES 2087–2098, DECEMBER 1988

(12)

Nonparametric Statistical Methods for Comparing Two Sites Based on Data With Multiple Nondetect Limits

Los Alamos Environmental Restoration Facility

STEVEN P. MILLARD

CH2M Hill, Bellevue, Washington

STEVEN J. DEVEREL

Water Resources Division, U.S. Geological Survey, Sacramento, California

ER Record I.D.# 0054953

As concern over the effects of trace amounts of pollutants has increased, so has the need for statistical methods that deal appropriately with data that include values reported as "less than" the detection limit. It has become increasingly common for water quality data to include censored values that reflect more than one detection limit for a single analyte. For such multiply censored data sets, standard statistical methods (for example, to compare analyte concentration in two areas) are not valid. In such cases, methods from the biostatistical field of survival analysis are applicable. Several common two-sample censored data rank tests are explained, and their behaviors are studied via a Monte Carlo simulation in which sample sizes and censoring mechanisms are varied under an assumed lognormal distribution. These tests are applied to shallow groundwater chemistry data from two sites in the San Joaquin Valley, California. The best overall test, in terms of maintained α level, is the normal scores test based on a permutation variance. In cases where the α level is maintained, however, the Peto-Prentice statistic based on an asymptotic variance performs as well or better.

INTRODUCTION

The continuing evolution of analytical chemistry techniques has brought about the ability to detect increasingly smaller concentrations of chemicals in the environment. Concomitantly, there is a growing public concern over the biological effects of chemicals in trace amounts. The justification of these concerns is reflected in the events at Kesterson National Wildlife Refuge, California in the early part of this decade [*Marshall*, 1985], where the bioaccumulation of Selenium has threatened wildlife. Measurements close to the limit of detection of an analytical technique, however, are usually extremely variable. It is therefore important to be able to use valid statistical techniques when describing trace chemical data.

The data in Table 1 are measurements of the concentration of copper and zinc (micrograms per liter) in shallow groundwater from two geological zones in the San Joaquin Valley, California [*Deverel et al.*, 1984; *Deverel and Millard*, 1988]. These data display a common feature of groundwater quality data: multiple detection limits. There are at least three possible causes of multiple detection limits. First, the limit of detection of a particular analyte depends upon the method that is used to measure it. There may be more than one method available, and each method may be optimal (have the smallest percent measurement error) in a certain range of analyte concentration. For example, the protocol may call for method 1 to be used if the specific conductance is above a certain threshold c and method 2 if specific conductance is below c.

A second cause of multiple detection limits involves the process of dilution. Due to time constraints, a lab technician may follow a protocol of allowing only a certain maximum number of dilutions for any single lab sample. Because the detection limit depends on the amount of dilution, multiple detection limits may result. If a study is conducted over a period of years, then a third cause of multiple detection limits may be decreasing detection limits over time as the measurement technique improves.

In this paper, data that display many detection limits, such as those of Table 1, will be denoted multiply censored, while data with only one detection limit will be termed singly censored. Standard parametric statistical methods such as t tests and multiple regression cannot be validly applied to singly or multiply censored data sets; it is not clear how the censored observations should be treated. Statisticians in the fields of survival analysis and life testing, however, have developed numerous techniques for analyzing multiply censored data sets [e.g., *Kalbfleisch and Prentice*, 1980]. Sometimes a specific parametric (e.g., exponential) model is assumed, allowing a maximum likelihood approach to estimation and testing. Water quality data, however, usually appear to follow nonstandard distributions and are often characterized by outliers and missing values. Thus nonparametric methods are commonly used to analyze water quality data [*Hipel*, 1988].

This paper discusses nonparametric approaches to comparing the concentration of a pollutant between two geographic areas based on multiply censored data. In statistical jargon, this is referred to as the two-sample location problem [e.g., *Hollander and Wolfe*, 1973, p. 67]. The term "location" refers to the location of central tendency (e.g., median or mean) of the probability distribution of the pollutant, not to a specific geographic area. The population upon which the probability distribution of the pollutant is based is the set of all possible measurements within a geographic area. The key question to be answered is, Does the location of central tendency differ between two geographic areas (i.e., is the median pollutant concentration the same in each area)?

First, previous work on censored data in the survival analysis and environmental monitoring literature is briefly reviewed. Next, standard nonparametric two-sample tests for uncensored or singly censored data are reviewed. The extension of these tests to multiply censored data is then given. A

Copyright 1988 by the American Geophysical Union.

Paper number 88WR03412.
0043-1397/88/88WR-03412$05.00

Received by ER-RPF
SEP 3 1996

TABLE 1. Groundwater Concentrations of Copper and Zinc at Two Geological Zones in the San Joaquin Valley, California

	Alluvial Fan Zone					Basin-Trough Zone					
Location	Cu	Zn	Location	Cu	Zn	Location	Cu	Zn	Location	Cu	Zn
1	<1	<10	36	20	10	1	2	20	36	4	30
2	<1	9	37	MS	20	2	2	10	37	8	25
3	3	MS	38	MS	620	3	12	60	38	1	<10
4	3	5	39	16	40	4	2	20	39	15	10
5	5	18	40	<5	50	5	1	12	40	3	40
6	1	<10	41	1	33	6	<10	8	41	3	20
7	4	12	42	2	10	7	<10	<10	42	1	10
8	4	10	43	<5	20	8	4	14	43	6	20
9	2	11	44	3	10	9	<10	<10	44	3	20
10	2	11	45	2	10	10	<1	17	45	6	30
11	1	19	46	8	10	11	1	<3	46	3	20
12	2	8	47	7	30	12	<2	11	47	4	30
13	<5	<3	48	5	20	13	<2	5	48	5	50
14	11	<10	49	<5	10	14	1	12	49	14	90
15	<1	<10	50	2	20	15	2	4	50	4	20
16	2	10	51	<10	20	16	<10	3			
17	2	10	52	<5	20	17	3	6			
18	2	10	53	<5	<10	18	<1	3			
19	2	10	54	2	20	19	1	15			
20	<20	<10	55	10	23	20	1	13			
21	2	10	56	2	17	21	3	4			
22	2	<10	57	4	10	22	<5	20			
23	3	10	58	<5	<10	23	MS	20			
24	3	<10	59	2	10	24	17	70			
25	MS	10	60	3	20	25	23	60			
26	<20	<10	61	9	29	26	9	40			
27	<10	10	62	<5	20	27	9	30			
28	7	10	63	2	<10	28	3	40			
29	5	20	64	2	10	29	3	17			
30	2	20	65	2	<10	30	<15	10			
31	2	<10	66	2	10	31	<5	20			
32	<10	20	67	1	7	32	4	20			
33	7	20	68	1	<10	33	<5	5			
34	12	20				34	<5	10			
35	<1	<10				35	<5	50			

Data from *Deverel et al.* [1984]. All concentrations are given in micrograms per liter. MS, missing value.

Monte Carlo simulation of these two-sample tests is discussed, and finally these tests are applied to the data of Table 1.

PREVIOUS STUDIES

Because survival analysis and life testing have laid the groundwork for censored data techniques, it is important to distinguish between the kinds of censoring that arise in these fields of study. One important distinction is between censoring and truncation. A sample is truncated on the left (right) if only values above (below) a known truncation point, say, t_0, are reported. The number of otherwise possible sample values excluded from the truncated sample is unknown [*Cohen*, 1959]. An example of a left-truncated sample is one in which only values above the detection limit are reported or used, and nondetects are excluded. For example, suppose the detection limit for a particular analyte at a particular laboratory is 5 ppb. Further, suppose that nine samples are analyzed with the results that six samples fall below the detection limit, and the other three samples yield values (in ppb) of 10, 15, and 20. If a data analyst then uses only the observations 10, 15, and 20 for a statistical analysis, and is unaware of or ignores the six other observations, then she/he is using a left-truncated sample.

A sample of n observations is singly censored on the left (right) if n_c ($n_c \geq 1$) of these observations are known only to fall below (above) a known censoring level, say, c, while the remaining n_u ($n_u = n - n_c$) uncensored observations falling above (below) c are fully reported. A sample of n observations is multiply censored with m censoring level if n_{c_1} observations are censored at censoring level c_1, n_{c_2} observations are censored at censoring level c_2, $\cdots$, n_{c_m} observations are censored at censoring level c_m, and uncensored observations fall in between each of the censoring levels. For example, the copper data for the alluvial fan zone in Table 1 represent multiply left-censored data with four censoring levels (1, 5, 10, and 20 μg/L).

A second important distinction is between type I and type II censoring. A singly censored sample of n observations arises from type I censoring on the left (right) if a specific censoring level c_1 is fixed in advance, and values below (above) c_1 are simply reported as less (greater) than c_1. A sample of n observations arises from type II censoring on the right if only the r ($1 \leq r < n$) smallest observations are fully reported, and the remaining n-r observations are known only to fall above the rth smallest value. Thus under type II censoring, values that appear extreme (larger than the rth-order statistic) relative to the rest of the observations are censored. Type II censoring can arise in life testing, where, for example, all n electronic components are started at the same time, and the experiment is stopped after r of the components have failed.

Environmental data sets involving detection limits almost

always fall into the category of type I left censoring. On the other hand, survival analysis studies almost always involve right-censored data. An example is the well-known Stanford Heart Transplant Study [*Crowley and Hu*, 1977], in which the (possibly right censored) survival times of potential heart transplant recipients were reported. Fortunately, statistical methods for right-censored data are easily modified to apply to left-censored data. A summary of statistical techniques from the survival analysis literature for type I censored data can be found in such texts as *Kalbfleisch and Prentice* [1980], *Lee* [1980], and *Miller* [1981].

In the environmental literature, *Gilbert and Kinnison* [1981], *Gilliom and Helsel* [1986], *Gleit* [1985], *Helsel and Gilliom* [1986], *Kushner* [1976], and *Owen and DeRouen* [1980] are all studies concerned with estimating parameters based on singly censored data sets. *Gilliom et al.* [1984] showed the effects of censoring with one detection limit on the power of Kendall's seasonal test for trend that was introduced by *Hirsch et al.* [1982]. There has been very little work in the field of water quality on techniques to handle multiply censored data sets.

NONPARAMETRIC TWO-SAMPLE LOCATION TESTS

The Wilcoxon Rank Sum test [*Hollander and Wolfe*, 1973, p. 67] is the standard nonparametric two-sample test for a difference in location of two distributions. It is equivalent to the Mann-Whitney U test [*Hollander and Wolfe*, 1973, p. 71] and hereafter abbreviated as the MWW test. The assumptions of the MWW test are that the probability distribution of the pollutant is the same in each area, except for a possible shift in median concentration. The null hypothesis is stated as there is no difference in median concentration between the two areas; the alternative hypothesis is there is a difference, or the median concentration in area 1 (area 2) is larger than the median concentration in area 2 (area 1).

In statistical notation, these hypotheses are written as

$$H_0: F_1(t) = F_2(t) \tag{1}$$

versus

$$H_a: F_1(t - \Delta) = F_2(t) \tag{2}$$

for all t, where F_i denotes the cumulative distribution function (cdf) of population i, $i = 1, 2$, and Δ denotes the shift in median concentration. Note that this kind of location shift is not applicable to cdf's that are bounded below or above by some constant. Thus in the case of two populations with lognormal distributions, a location shift hypothesis of the form (1) versus (2) is not applicable to the cdf's of the original observations, but is applicable to the cdf's of the log-transformed observations.

The null and one-sided alternative hypotheses can alternatively be written as

$$H_0: \Delta = 0 \tag{3}$$

versus

$$H_a: \Delta > 0 \tag{4}$$

If $\Delta > 0$, population 2 has a larger median concentration than population 1. The other one-sided alternative ($\Delta < 0$) and the two-sided alternative ($\Delta \neq 0$) could be considered as well.

The assumptions of the MWW test imply that the distribution of the pollutant does not differ in variability between sites, even if the medians do. The MWW test, however, usually works quite well for testing the more general alternative hypothesis

$$H_a: F_1(t) \geq F_2(t) \tag{5}$$

with strict inequality for at least one t. Note that the hypotheses (1) versus (5) allow for cdf's that are not bounded below by some constant.

To apply the MWW test, the following notation is used. Assume $X_1, \cdots, X_{m_1}$ are independent and identically distributed (iid) observations from population 1 with cdf F_1, and $Y_1, \cdots, Y_{m_2}$ are iid observations with cdf F_2. Let $N = m_1 + m_2$. The X and Y are combined and ranked. The MWW statistic W is computed as

$$W = \sum_{i=1}^{m_2} R_i \tag{6}$$

where $R_1, \cdots, R_{m_2}$ denote the ranks of the Y in the combined sample. The null hypothesis (1) is rejected in favor of the alternative (2) or (5) if W is too large. For small sample sizes, the exact distribution of W under H_0 may be found in tables [e.g., *Hollander and Wolfe*, 1973, pp. 272–282]. For larger samples, a normal approximation is used [*Hollander and Wolfe*, 1973, p. 68].

SCORE TESTS AND LOCALLY MOST POWERFUL RANK TESTS

There are several other nonparametric two-sample location tests that employ statistics of a form similar to (6). They can be written generally as

$$L = \sum_{i=1}^{m_2} a(R_i) \tag{7}$$

where $a(\ )$ is called the score function because L is a score statistic (see, for example, *Cox and Hinkley* [1974] for an explanation of score tests and statistics). Statistics of the form (7) are also called linear rank statistics [*Prentice*, 1985]. The MWW test is a linear rank statistic that uses the score function $a(R_i) = R_i$.

Under the null hypothesis (1), the distribution of the linear rank statistic L in (7) does not depend upon the form of the underlying distribution F_1. Hence tests based on L are called nonparametric or distribution-free. If H_0 is not true, however, then the distribution of L will depend not only upon the distributions F_1 and F_2, but also upon the form of the score function $a(\ )$.

The decision of what scores to use may be based on consideration of the power of the test. A rank test of the hypotheses (1) versus (2) is a locally most powerful rank test (LMPRT) of size α if it maximizes the slope of the power function (as a function of Δ) at $\Delta = 0$, among all possible size α rank tests [*Hettmansperger*, 1984, p. 144]. Figure 1 illustrates the idea of a locally most powerful test. Of the three tests shown, test A is the locally most powerful one. A LMPRT yields the maximum power in the neighborhood of $\Delta = 0$, i.e., for small deviations from H_0. Thus if the LMPRT can be written in the form of (7), one obvious choice for the score function is the one that yields the LMPRT.

Hettmansperger [1984, pp. 144–145] shows that the statistic associated with the LMPRT of the hypotheses (1) versus (2) is the score statistic of (7) with scores

$$a(i) = -E[f_1'(V_{(i)})/f_1(V_{(i)})] \tag{8}$$

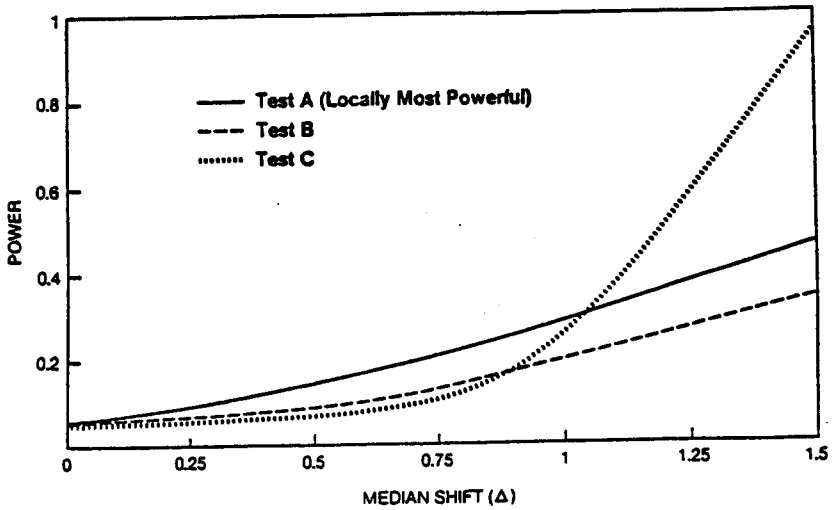

Fig. 1. Illustration of a locally most powerful test.

where f_1 denotes the probability density function (pdf) of population 1, $f_1'(x)$ denotes the derivative of f_1 at x, and $V_{(1)}, \cdots, V_{(N)}$ denote the order statistics from a sample of size N with pdf f_1 (or cdf F_1). One can generate the scores associated with LMPRT's by substituting in various pdf's in (8). Table 2 gives the scores associated with some common LMPRT's.

Because the test based on the score $a(R_i) = R_i$ is equivalent to the one based on the score shown in Table 2 for the logistic distribution, the MWW test is the LMPRT when the observations follow this distribution. Similarly, since the ranks based on observations from a lognormal distribution are the same as those based on the log-transformed observations, the normal scores test is the LMPRT when the underlying distribution is either normal or lognormal. (In the case of a lognormal distribution, the LMPRT refers to a test for the difference between the medians of the log-transformed observations.) A large sample normal approximation to the distribution of L is given by *Hettmansperger* [1984, p. 148].

SINGLY CENSORED DATA

The extension of the MWW test to singly censored data (i.e., data containing observations censored at one detection limit) is straightforward, assuming there are no uncensored observations with values below the detection limit. All observations below the detection limit are considered to be tied when the ranks are assigned. There are several possible ways to handle tied observations [*Pratt and Gibbons*, 1981] including (1) assign the average rank to each tied observation, (2) assign the lower ranks to the Y observations in the tied group, (3) assign the higher ranks to the Y observations in the tied group, and (4) randomly assign distinct ranks to the tied observations. Table 3 demonstrates the ranks and W statistics associated with these four methods for a hypothetical data set.

For methods 2–4, the usual MWW statistic is compared to its critical value that is computed as if there were no ties. Method 2 is conservative (the true α level is below the nominal α level), method 3 is liberal, and method 4 preserves the α level under H_0. The most commonly used method, however, is method 1. When average ranks are assigned to tied observations, the standard tables of the permutation distribution of W are not valid [*Lehmann*, 1975, p. 18]; thus the permutation distribution must be explicitly derived, or a modification of the large-sample approximation must be used. For the large-sample normal approximation, the variance of W is modified to account for ties, and is smaller than if ties were not present [*Hollander and Wolfe*, 1973, p. 69]. Note that uncensored observations may be tied as well.

Although methods 1 and 4 both preserve the α level, method 4 suffers from arbitrariness (two different researchers can produce two different answers). Also, method 1 is more powerful than method 4 [*Putter*, 1955; *Buhler*, 1967]. In general, the best way to extend linear rank tests of the form (7) to

TABLE 2. Scores of LMPRT's for Various Distributions

Distribution	Score, $a(R_i)$	Test Name
Logistic	$[2/(N+1)]R_i - 1$	Wilcoxon rank sum
Normal or lognormal	$\Phi^{-1}[R_i/(N+1)]$*	Van der Waerden or normal scores
Exponential or extreme value	$\sum_{j=1}^{R_i} (N-j+1)^{-1}$	Savage scores
Double exponential	$\mathrm{sgn}[R_i - (N+1)/2]$*	Mood's median test

The function $a(\ )$ is given in (8). The Φ denotes the cdf of the standard normal distribution. Here $\mathrm{sgn}(x) = 1$ if $x > 0$, $\mathrm{sgn}(x) = 0$ if $x = 0$, and $\mathrm{sgn}(x) = 1$ if $x < 0$.
*Denotes an approximation of the true score.

TABLE 3. Four Methods for Computing the MWW W statistic in the Presence of Ties

Method	Rank Vector	W
Average ranks	(2, 2, 4, 2, 5)	7
Lower ranks	(2, 3, 4, 1, 5)	6
Higher ranks	(1, 2, 4, 3, 5)	8
Random ranks	Randomly choose	
	(1, 2, 4, 3, 5)	8
	(1, 3, 4, 2, 5)	7
	(2, 1, 4, 3, 5)	8
	(2, 3, 4, 1, 5)	6
	(3, 1, 4, 2, 5)	7
	(3, 2, 4, 1, 5)	6

Data: $(X_1, X_2, X_3, Y_1, Y_2) = (7^-, 7^-, 8, 7^-, 9)$; 7^- denotes an observation (left−) censored at 7.

TABLE 4. Censored Data Notation Applied to the Copper Data of Table 1

i	t_i	n_{Li}	n_{L1i}	n_{L2i}	d_i	d_{1i}	d_{2i}	e_i	e_{1i}	e_{2i}
1	1	18	9	9	12	5	7	6	4	2
2	2	45	30	15	25	21	4	2	0	2
3	3	59	36	23	14	6	8	0	0	0
4	4	67	39	28	8	3	5	0	0	0
5	5	84	50	34	4	3	1	13	8	5
6	6	86	50	36	2	0	2	0	0	0
7	7	89	53	36	3	3	0	0	0	0
8	8	91	54	37	2	1	1	0	0	0
9	9	94	55	39	3	1	2	0	0	0
10	10	102	59	43	1	1	0	7	3	4
11	11	103	60	43	1	1	0	0	0	0
12	12	105	61	44	2	1	1	0	0	0
13	14	106	61	45	1	0	1	0	0	0
14	15	108	61	47	1	0	1	1	0	1
15	16	109	62	47	1	1	0	0	0	0
16	17	110	62	48	1	0	1	0	0	0
17	20	113	65	48	1	1	0	2	2	0
18	23	114	65	49	1	0	1	0	0	0

the case of singly censored and/or tied observations is to compute the scores as if there were no ties and then average the scores of the tied observations.

MULTIPLY CENSORED DATA

The extension of two-sample rank tests for shifts in location to multiply censored data has been studied by several authors, including *Breslow* [1970], *Cox* [1972], *Gehan* [1965], *Mantel* [1966], *Peto and Peto* [1972], and *Prentice* [1978]. All of these authors were concerned primarily with right-censored data, but their methods are also applicable to left-censored data. For the case of uncensored data, it was shown that a score test based on the score statistic (7) is "optimal" (it is the LMPRT) for testing hypotheses (1) versus (2). Prentice demonstrated that all the tests for multiply censored data proposed by the above authors could be written as score tests as well, based on a particular definition of the probability distribution of the rank vector. In order to explain the reasoning behind these tests, it will be helpful to introduce some notation that closely follows that of Prentice, *Prentice and Marek* [1979], and *Latta* [1981].

As before, let m_1 and m_2 denote the number of observations in samples 1 and 2, respectively, and set $N = m_1 + m_2$. Let $t_1 < t_2 < \cdots < t_k$ denote the ordered distinct uncensored observations for the combined samples. (In the context of survival data, t stands for the "time" of death.) For $i = 1, \cdots, k$, let d_{1i} denote the number of observations from sample 1 that are equal to t_i, and similarly for d_{2i}. Set $d_i = d_{1i} + d_{2i}$ (the number of "deaths" at time t_i). If there are no tied uncensored observations, then d_i will always be equal to 1.

For right-censored data, e_{1i} equals the number of censored observations from sample 1 with censoring levels that fall into the interval $[t_i, t_{i+1})$, where $t_{k+1} = +\infty$; similarly, for e_{2i}. Set $e_i = e_{1i} + e_{2i}$. Let n_{R1i} be the number of observations from sample 1 known to be at least as large as t_i; similarly, for n_{R2i}. Set $n_{Ri} = n_{R1i} + n_{R2i}$.

For left-censored data, e_{1i} equals the number of censored observations from sample 1 with censoring levels that fall into the interval $(t_{i-1}, t_i]$, where $t_0 = -\infty$; similarly for e_{2i}. Let n_{L1i} be the number of observations from sample 1 known to be less than or equal to t_i; similarly, for n_{L2i}. Define e_i and n_{Li} as for right-censored data. An illustration of this notation for the left-censored copper data of Table 1 is given in Table 4.

For the two-sample case, Prentice's score statistic v can be written in the form

$$v = \sum_{i=1}^{k} (d_{2i}c_i + e_{2i}C_i) \tag{9}$$

where c_i and C_i denote the scores associated with the uncensored and censored observations, respectively. In the case of the score statistic (7) for uncensored observations, several different candidate score functions $a(\ )$ were derived based on using (8) and assuming various distributions for F_1 (see Table 2). In the case of multiply censored data, the equations analogous to (8) for computing the proper scores c_i and C_i associated with various distributions for F_1 are given in (17) of *Prentice* [1978], and approximations to these scores are given in (30) of the same paper.

Table 5 lists Prentice's scores for some assumed distributions, for both the left- and right-censored case. These scores are written so that as before, if the observations in the second sample tend to be larger than those in the first, the statistic will be large. In the case of no censoring, the Peto-Prentice statistic reduces to one that is equivalent to the

TABLE 5. Scores of Some Censored Data Rank Tests

Distribution	Left Censored c_i	Left Censored C_i	Right Censored c_i	Right Censored C_i	Test Name	Uncensored Analogue
Logistic	$2\hat{F}_{Li} - 1$ $n_{Li} - (k-i+1)$	$\hat{F}_{Li} - 1$ $-(k-i+1)$	$2\hat{F}_{Ri} - 1$ $i - n_{Ri}$	$\hat{F}_{Ri}$ i	Peto-Prentice Gehan or Breslow	Wilcoxon rank sum Wilcoxon rank sum
Normal or lognormal	$\Phi^{-1}(\hat{F}_{Li})$	$-(1/\hat{F}_{Li})\phi(c_i)$ or (10)	$\Phi^{-1}(\hat{F}_{Ri})$	$(1/\hat{S}_{Ri})\phi(c_i)$ or (10)	normal scores	Van der Waerden or normal scores
Exponential or extreme value	$1 - \sum_{j=i}^{k} n_{Lj}^{-1}$	$-\sum_{j=i}^{k} n_{Lj}^{-1}$	$\sum_{j=1}^{i} n_{Rj}^{-1} - 1$	$\sum_{j=1}^{i} n_{Rj}^{-1}$	log rank	Savage scores
Double exponential	$\mathrm{sgn}(\hat{F}_{Li} - 0.5)$	*	$\mathrm{sgn}(\hat{F}_{Ri} - 0.5)$	*	generalized sign test	Mood's median test

An entry under Distribution denotes a score derived via *Prentice's* [1978] method of scoring; ϕ denotes the pdf of the standard normal distribution. $\hat{F}_{Li} = \Pi_{j=i}^{k}[n_{Lj} - d_j + 1)/(n_{Lj} + 1)]$; $\hat{S}_{Li} = 1 - \hat{F}_{Li}$. $\hat{F}_{Ri} = 1 - \hat{S}_{Ri}$; $\hat{S}_{Ri} = \Pi_{j=1}^{i}[(n_{Rj} - d_j + 1)/(n_{Rj} + 1)]$.

*For left-censored data, $C_i = -\hat{S}_{Li}/(1 - \hat{S}_{Li})$ if $\hat{S}_{Li} < 0.5$, and $C_i = -1$ if $\hat{S}_{Li} \geq 0.5$. For right-censored data, $C_i = \hat{F}_{Ri}/(1 - \hat{F}_{Ri})$ if $\hat{F}_{Ri} < 0.5$, and $C_i = 1$ if $\hat{F}_{Ri} \geq 0.5$. See text for further explanation.

TABLE 6. The 12 Censored Data Rank Tests Used in the Monte Carlo Study

Number	Name	Explanation
1	GP	Gehan, permutation variance.
2	LP	log rank, permutation variance.
3	PPP	Peto-Prentice, permutation variance.
4	GH	Gehan, hypergeometric variance.
5	LH	log rank, hypergeometric variance.
6	PPH	Peto-Prentice, hypergeometric variance.
7	PPA	Peto-Prentice, asymptotic variance.
8	MWW1	MWW with all observations less than largest censoring level considered to be tied
9	MWW2	MWW with all censored observations set equal to half the censoring level prior to ranking
10	NS1P	normal scores, permutation variance. C_i based on (30) of *Prentice* [1978]
11	NS2P	normal scores, permutation variance. C_i based on (10).
12	NS2H	normal scores, hypergeometric variance. C_i based on (10).

MWW rank sum statistic. *Gehan* [1965] and *Breslow* [1970] also developed an extension to the MWW statistic for censored data, but not in the context of *Prentice*'s [1978] method of scoring. Their statistic, however, can still be written in the form (9), and the associated scores are also shown in Table 5.

The quantities $\hat{F}_i$ and $\hat{S}_i$ in Table 5 are estimates of the cdf and survival function, respectively, at t_i for the combined groups. If $F(x)$ denotes the cdf of a random variable X, then the survival function is given by $S(x) = Pr(X > x) = 1 - F(x)$. This notation is the only major deviation from the survival analysis literature, where F is often used to denote the survival function rather than the cdf.

If ties occur among the uncensored observations, then, as for the case of singly censored data, the scores should be calculated as if there were no ties, and then the average scores should be assigned to the tied observations [*Prentice*, 1978; *Latta*, 1981].

The Peto-Prentice, Gehan, and logrank statistics are fairly common in the survival analysis literature [e.g., *Prentice and Marek*, 1979]. The normal scores statistic has not been used as often, although it would appear to be better suited for water quality data, since these data often exhibit positively skewed distributions that appear close to being lognormally distributed. The normal scores c_i shown in Table 5 were derived from the approximation (30) of *Prentice* [1978]. There are two distinct methods of deriving the normal scores C_i: (1) from approximation (30) of Prentice or (2) by using the following relation given in *Prentice and Marek* [1979]:

$$C_i = (n_i C_{i-1} - c_i)/(n_i - 1) \tag{10}$$

for $i = 1, \cdots, k$, where $C_0 = 0$, and n_i denotes n_{Li} or n_{Ri}, depending on whether the data are left- or right-censored. For left-censored data, C_1 is not defined if $n_{L1} = 1$, nor is C_k defined for right-censored data if $n_{Rk} = 1$. In each case these scores may be set arbitrarily to 0, since $e_{21} = 0$ or $e_{2k} = 0$, respectively, so that neither score will contribute to the computation of v or, as we shall see, the variance of v.

Under the null hypothesis (1), the expected value of v in (9) is 0. In order to test the hypotheses (1) versus (5), v must be computed and divided by an estimate of its standard deviation. This standardized z statistic will be approximately distributed as a standard normal random variable under the null hypothesis (1) [*Prentice*, 1978]. *Latta* [1981] describes three ways to derive an estimate of the variance of v. If the censoring mechanism is the same for both samples, then the variance is derived by assuming each of the $\binom{N}{m_1}$ possible divisions of observations into two samples of size m_1 and m_2 is equally likely. This permutation variance estimate is given by

$$\mathrm{Var}_P = \{m_1 m_2/[N(N-1)]\} \sum_{i=1}^{k} (d_i c_i^2 + e_i C_i^2) \tag{11}$$

If the censoring mechanism differs between the two groups, then the permutation variance is not appropriate, and a conditional permutation approach must be used. In this case, *Prentice and Marek* [1979] rewrite v as

$$v = \sum_{i=1}^{k} w_i[d_{2i} - d_i(n_{2i}/n_i)] \tag{12}$$

where $w_i = c_i - C_i$ and n_i and n_{2i} are determined appropriately depending on whether the data are left- or right-censored. (See Table 5 to compute specific values of w_i.) The conditional permutation or hypergeometric variance estimate of v (12) is given by

$$\mathrm{Var}_H = \sum_{i=1}^{k} d_i w_i^2 (n_{2i}/n_i)[1 - (n_{2i}/n_i)][(n_i - d_i)/(n_i - 1)] \tag{13}$$

For left-censored data, the first term in the sum is taken to be 0 if $n_{L1} = 1$, and for right-censored data the final term is 0 if $n_{Rk} = 1$.

A third variance estimator, derived by *Prentice* [1978] and based on the log likelihood of the rank vector, is the asymptotic variance estimator, Var_A. Var_A is the same as Var_H for the logrank and Gehan statistics (assuming no ties in the uncensored observations), but for the Peto-Prentice statistic, for right-censored data, it is given by

$$\mathrm{Var}_A = \sum_{i=1}^{k} \left\{[\hat{S}_{Ri}(1 - a_i)b_i - (a_i - \hat{S}_{Ri})b_i] \cdot \left[\hat{S}_{Ri} b_i + 2 \sum_{j=i+1}^{k} \hat{S}_{Rj} b_j\right]\right\} \tag{14}$$

where $a_i = \pi_{j=1}^{i}(n_{Rj} + 1)/(n_{Rj} + 2)$ and $b_i = 2d_{2i} + e_{2i}$. In the case of tied uncensored observations, the expressions (12) and (13) do not need to be modified; their form allows for ties. For the asymptotic variance estimator Var_A of (14), however, the scores a_i and $\hat{S}_{Ri}$ should be computed as if there were no ties, and then the average scores assigned to the tied observations [*Prentice*, 1978; *Latta*, 1981].

TABLE 7. The Nine Conditions of Sample Sizes and Censoring for the Monte Carlo Study

	Sample Size		
	Area 1 = Area 2	Area 1 > Area 2	Area 1 < Area 2
Area 1 = area 2	1	2	3
Area 1 > area 2	4	5	6
Area 1 < area 2	7	8	9

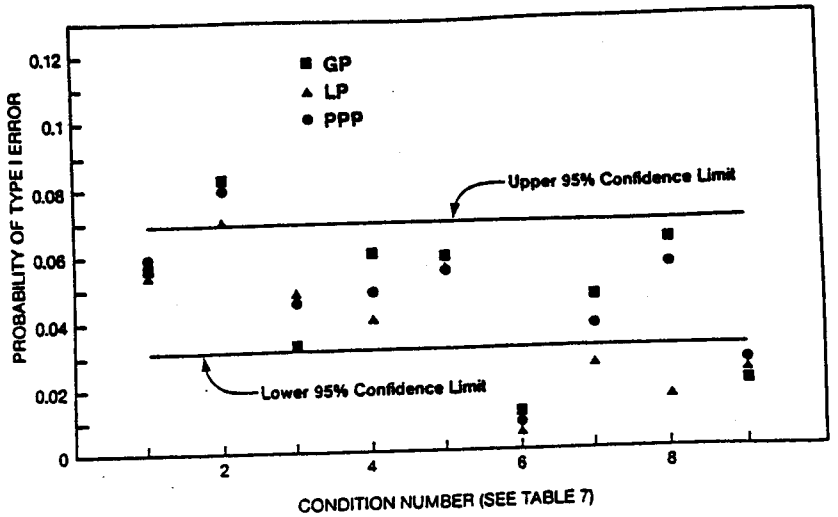

Fig. 2. Empirical α levels of the GP, LP, and PPP tests.

A MONTE CARLO STUDY

A study of the empirical α levels and powers of some two-sample censored data rank tests was carried out for various conditions of sample sizes and censoring using an AT-compatible personal computer and the statistical package GAUSS. Table 6 lists the 12 tests that were studied. Test 8 is the MWW test applied to data in which all observations (censored or uncensored) below the largest detection limit are treated as tied. This approach was suggested by *Hirsch et al.* [1982] and *Gilliom et al.* [1984] in the context of testing for trend. Test 9 is the MWW test applied to data for which censored observations are set equal to half the value of the detection limit before being ranked. This approach has been used by *Nehls and Akland* [1973], *Gleit* [1985], *Deverel and Millard* [1988], and *Gilliom and Helsel* [1986].

In order to describe how the censoring mechanism was simulated, it will be helpful to introduce the following notation. Following *Miller* [1981], let $T_1, \cdots, T_{m_1}$ denote the true pollutant concentrations for the sample from area 1 and assume these are iid random variables with cdf F_1. If there were no censoring, we would observe $T_1, \cdots, T_{m_1}$. In the case of left censoring, however, what are actually observed are

$$X_i = \max(T_i, \tau_i) \tag{15}$$

where τ_i is the censoring level associated with observation i, $i = 1, \cdots, m_1$, and $\tau_1, \cdots, \tau_{m_1}$ are assumed to be iid random variables with cdf G_1. Similarly, if $U_1, \cdots, U_{m_2}$ denote the true pollutant concentrations for the sample from area 2, with cdf F_2, what are actually observed are

$$Y_j = \max(U_j, \mu_j) \tag{16}$$

where μ_j is the censoring level associated with observation j, $j = 1, \cdots, m_2$, and $\mu_1, \cdots, \mu_{m_2}$ are iid random variables with cdf G_2.

As an example of applying the above notation, suppose that four samples were taken in area 1 and analyzed, with each analysis subject to a detection limit of 5 ppb. Then τ_1, τ_2, τ_3, and τ_4 are iid random variables from a discrete uniform distribution with a probability mass of 1 at 5 ppb and a probability mass of 0 everywhere else. The cdf G_1 is written as

$$\begin{aligned} G_1(t) &= 0 \qquad t < 5 \\ G_1(t) &= 1 \qquad t \geq 5 \end{aligned} \tag{17}$$

If the analyses yielded (10, <5, <5, 8), then $(X_1, X_2, X_3, X_4) = (T_1, \tau_2, \tau_3, T_4)$.

For the Monte Carlo simulation, it was assumed that F_1 was the cdf of a lognormal distribution with mean and coef-

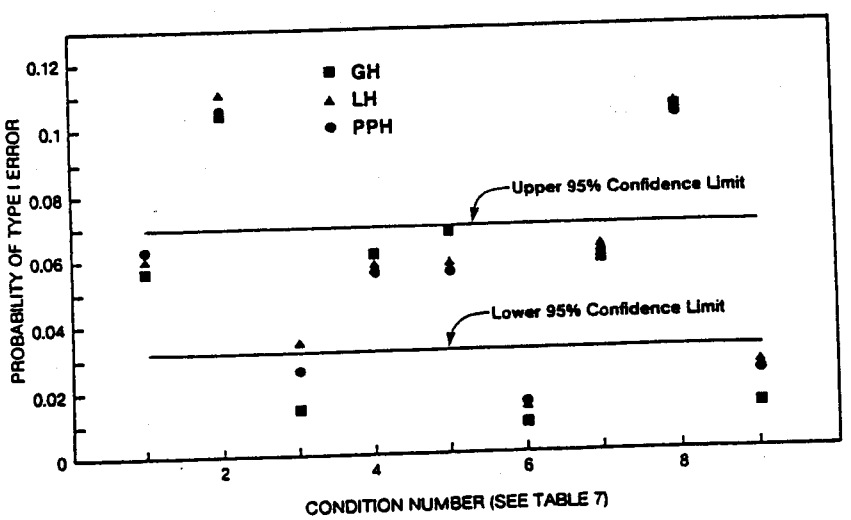

Fig. 3. Empirical α levels of the GH, LH, and PPH tests.

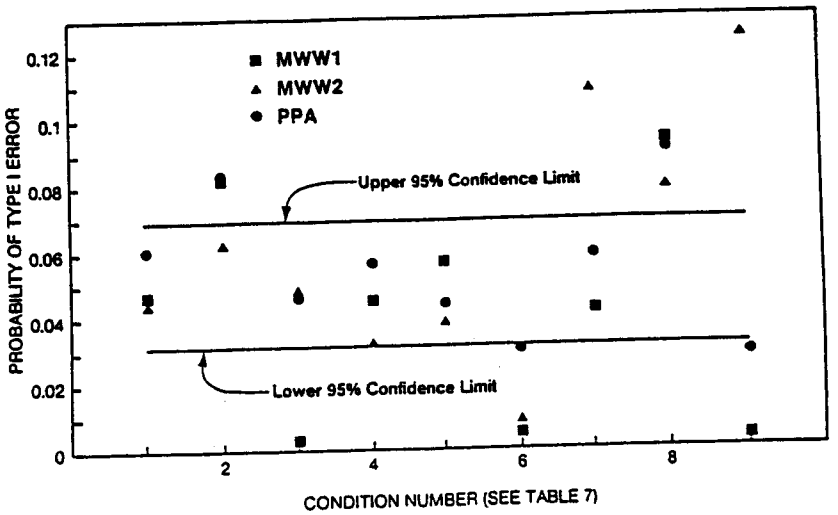

Fig. 4. Empirical α levels of the MWW1, MWW2, and PPA tests.

ficient of variation equal to 1. The pollutant concentrations from area 2 were also assumed to follow a lognormal distribution with a cv of 1, but showed various positive shifts in the median.

For this study, the distributions of the censoring levels were assumed to be discrete uniform (DU) and allowed for 2 or 4 censoring levels. The censoring levels used were the 20th, 40th, 60th, and 80th percentiles of F_1 (the distribution for area 1). Three patterns of censoring were considered:

(1) the same censoring mechanism for both areas,

$$G_1 = G_2 = DU\{q_{0.2}, q_{0.4}, q_{0.6}, q_{0.8}\}$$

(2) a larger proportion censored in the first area,

$$G_1 = DU\{q_{0.6}, q_{0.8}\} \qquad G_2 = DU\{q_{0.2}, q_{0.4}\}$$

(3) a larger proportion censored in the second area,

$$G_1 = DU\{q_{0.2}, q_{0.4}\} \text{ and } G_2 = DU\{q_{0.6}, q_{0.8}\}$$

where q_p denotes the pth percentile of F_1.

Censoring pattern 2 could arise, for example, in the case where areas 1 and 2 are in fact the same area, but samples from "area 2" represent those taken at a later time than the first batch of samples taken in area 1. If the analytical method has improved between the two time periods in which samples were taken, then the samples from area 1 would be subject to higher detection limits than the samples from area 2.

In addition to varying the censoring patterns, three combinations of sample sizes were considered: (1) equal sample sizes, $m_1 = m_2 = 10$; (2) sample 1 larger than sample 2, $m_1 = 20$, $m_2 = 5$; and (3) sample 1 smaller than sample 2, $m_1 = 5$, $m_2 = 20$. Table 7 summarizes the 9 conditions considered in the study.

For each condition, 500 trials were run. Each trial consisted of generating pollutant concentrations in each area, censoring the concentrations that fell below the detection limits, computing the 12 test statistics, and determining, for each test, whether the null hypothesis (1) was rejected.

Figures 2–5 illustrate the empirical α levels of the 12 tests for the nine conditions considered. In each case, the nominal α level was 0.05. (Results for the α levels 0.01, 0.025, and 0.10 showed similar patterns.) The solid horizontal lines in these figures represent the 95% confidence limits for the estimated α, assuming the true α level is 0.05. The upper and lower $(1 - p)$ 100% confidence limits are calculated as

$$CL = 0.05 \pm z_{1-(p/2)}[(0.05)(0.95)/500]^{1/2} \tag{18}$$

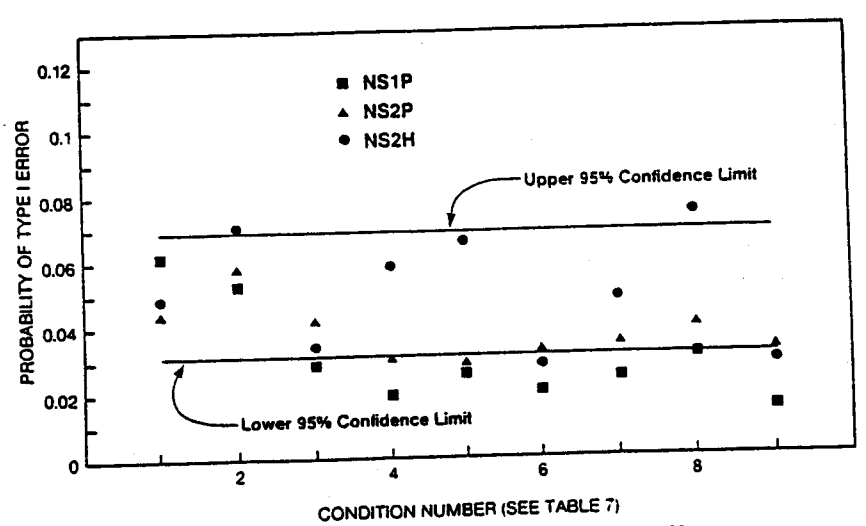

Fig. 5. Empirical α levels of the NS1P, NS2P, and NS2H tests.

TABLE 8. Number of Times the Empirical α Level Fell Outside the 95 and 99% Confidence Limits for the 9 Monte Carlo Conditions

Test Name	No. of Times Outside 95% CL	No. of Times Outside 99% CL
GP	3	3
LP	5	3
PPP	3	2
GH	5	5
LH	4	3
PPH	5	4
PPA	4	2
MWW1	5	5
MWW2	4	4
NS1P	7	4
NS2P	2	0
NS2H	4	0

where z_p denotes the pth percentile of the standard normal distribution.

The best behaved tests, in terms of maintaining the nominal α level over all nine conditions, were the normal scores tests with censored observation scores based on (10) (tests 11 and 12 of Table 6, denoted NS2P and NS2H). Table 8 shows how many times each of the tests fell outside the 95 and 99% confidence limits. Although the NS2H test fell outside the 95% confidence limits more than other tests, unlike the other tests both it and the NS2P test never fell outside of the 99% confidence limits.

The true α levels of the other tests were greatly affected by unequal sample sizes and unequal censoring mechanisms. In general, fewer observations in the shifted group combined with heavier censoring tended to increase the true α level, while more observations in the shifted group combined with less censoring tended to reduce it. The test that performed the poorest was the NS1P test (test 10).

A comparison of the powers of competing tests is not valid if the α levels of the tests are not comparable. Figures 6 and 7 illustrate the powers of the PPA, NS2P, and MWW1 tests for two conditions (1 and 6) in which the true α level was close to the nominal 0.05 α level for all three tests. In both cases, the PPA tests is as powerful of more powerful than the NS2P text. Not surprisingly, the power of the MWW1 test can fall far below that of the other two tests.

EXAMPLE

Deverel et al. [1984] and *Deverel and Millard* [1988] studied the distribution of trace elements in the groundwater in part of the San Joaquin Valley, California. The study area was divided into two geologic zones: the alluvial fan zone, consisting of material eroded from the Coast Range (to the west), and the basin trough zone, consisting of a mixture of Coast Range and Sierra Nevada (to the east) alluvium. As part of their analysis, *Deverel and Millard* [1988] compared trace element concentrations between these two zones using the MWW2 test previously described. The copper and zinc data will be reanalyzed in this paper.

Figures 8 and 9 show the estimated cdf's of copper and zinc concentrations for the two geological zones (see Table 5 for the formula for the estimated cdf's; this formula was applied to the data from each zone separately). Based on these figures, there appears to be no difference in copper concentration between zones, while there may be a higher concentration of zinc in the basin. Table 9 shows the values of the 12 statistics of Table 6 applied to these data, along with two-sided p values. None of the p values is significant for the copper concentrations, whereas all the p values are significant at the 0.10 level for zinc. Because these data show fairly equal sample sizes and censoring patterns similar to condition 1 of the Monte Carlo study, no major discrepancies between test results was expected.

DISCUSSION

Latta [1981] performed a similar but more extensive Monte Carlo power study of two-sample censored data rank tests. He used lognormal, Weibull, and exponential parent distributions (F_1 and F_2), but used a continuous uniform distribution for the censoring distributions (G_1 and G_2). He only considered the behaviors of tests 1–7 in Table 6, however. For a lognormal parent distribution, one would expect the normal scores tests (tests 10–12) to perform best, since these are the LMPRT's in the case of uncensored data. In the present study, two normal scores tests behaved best in terms of maintaining the nominal α level over several conditions, but they were not

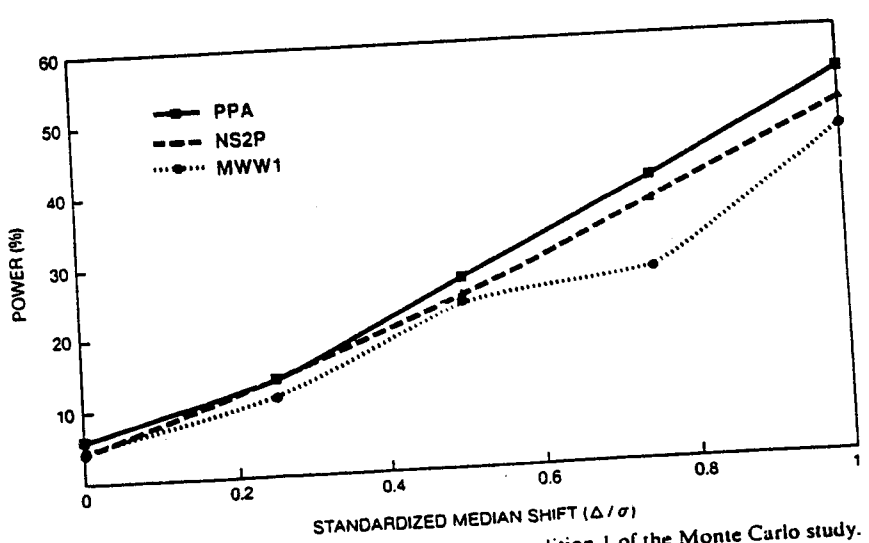

Fig. 6. Power curves for sample size and censoring condition 1 of the Monte Carlo study.

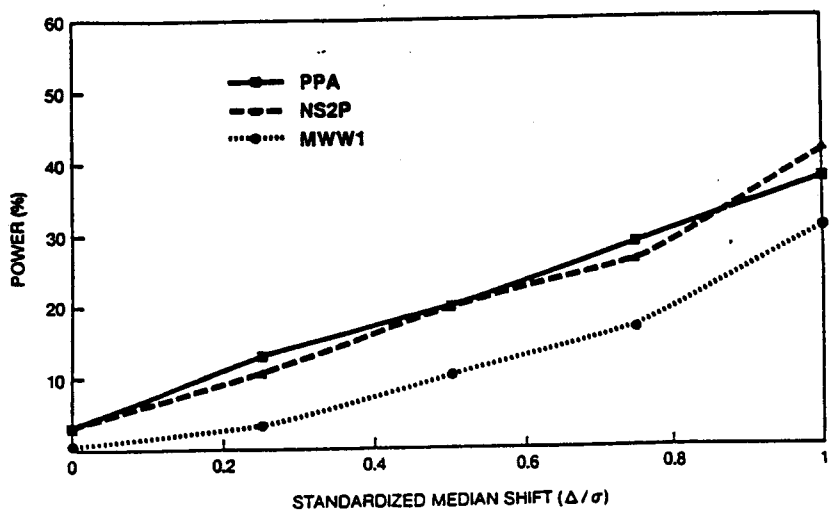

Fig. 7. Power curves for sample size and censoring condition 6 of the Monte Carlo study.

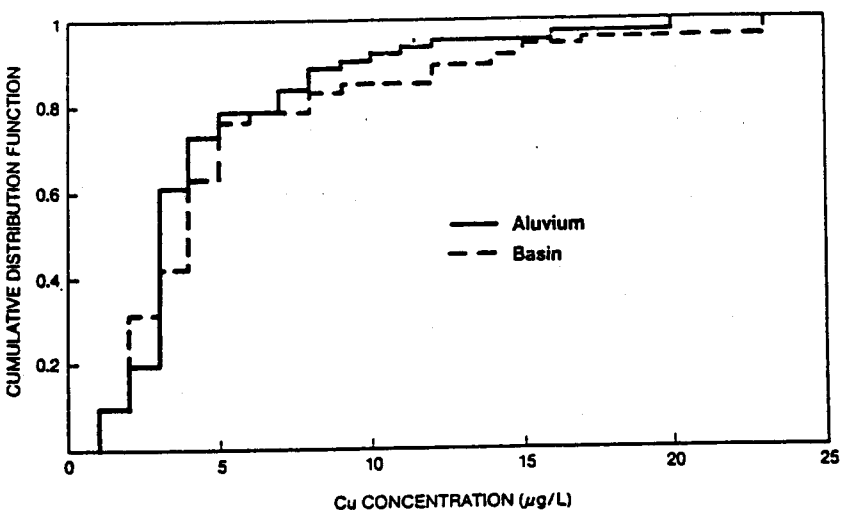

Fig. 8. Estimated cdf's of copper concentration in the two geological zones.

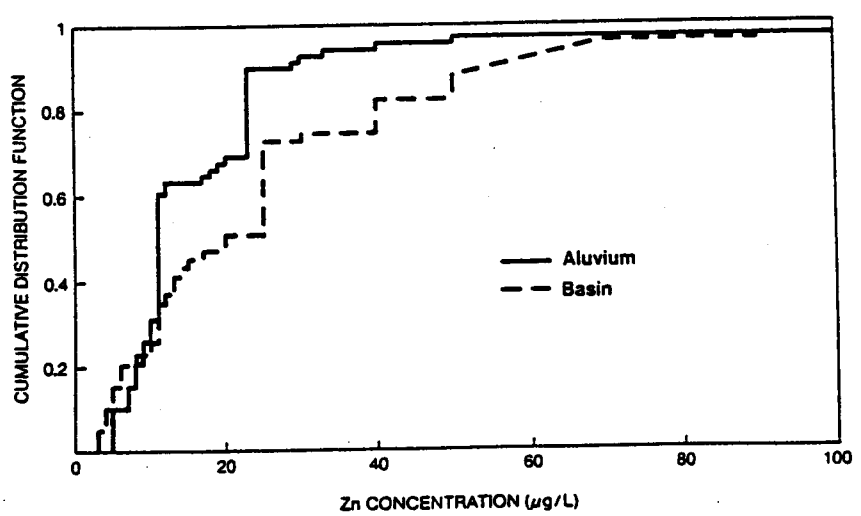

Fig. 9. Estimated cdf's of zinc concentration in the two geological zones.

TABLE 9. Results of the Two-Sample Censored Data Rank Tests of Table 6 Applied to the Groundwater Data of Table 1

	Cu		Zn	
Test	z	p	z	p
GP	0.8	0.42	2.49	0.01
LP	0.8	0.42	1.75	0.08
PPP	0.9	0.36	2.42	0.01
GH	0.7	0.48	2.35	0.02
LH	0.5	0.62	1.69	0.09
PPH	1.0	0.32	2.59	0.01
PPA	0.9	0.36	2.37	0.02
MWW1	0.2	0.84	2.36	0.02
MWW2	1.0	0.32	2.14	0.03
NS1P	0.9	0.36	2.37	0.02
NS2P	1.0	0.32	2.39	0.02
NS2H	0.9	0.36	2.36	0.02

Reported p values are two sided.

always the most powerful tests in cases where other tests also maintained the nominal α level. Indeed, *Latta* [1981] concluded from his study that the PPA test performed the best overall. As *Kalbfleisch and Prentice* [1980] and Latta point out, however, the efficiency (power) characteristics of censored data rank tests are not well-known.

For left-censored data, rather than using the scores of Table 5, one could just as well apply the scores appropriate for right-censored to left-censored data that is multiplied by -1, and then change the sign of the resulting z statistic. Hence existing software for analyzing right-censored data can be easily applied to left-censored data.

One strength of *Prentice*'s [1978] approach to censored data rank tests is that it was developed in the context of a general linear model. Hence his method of scoring can be used, for example, to develop tests for trend or to compare k samples ($k > 2$).

Another possible approach to comparing samples with multiply censored data is to combine the data and assign the average rank, where the average is taken over all permissible permutations of the rank vector. Once these ranks are obtained, they may be used with any of the standard linear rank tests described in Table 2. Such an approach was used by *Hughes and Millard* [1988] to extend Kendall's seasonal tau [*Hirsch et al.*, 1982] to the case of multiply censored data. For the two-sample case, it turns out that this approach applied to the MWW test yields Gehan's test with the permutation variance estimator (test 1 of Table 6).

Although this paper has demonstrated several possible tests for the two-sample location problem with multiply censored data, many further areas of research remain. One is the problem of simulating censoring mechanisms that actually occur in practice. Modeling these mechanisms will of course depend on the causes of the multiple detections limits. A second problem to be considered is the behavior and loss of power of these tests in the presence of ties in the uncensored observations. Finally, there is the question of how these tests behave in the presence of heterogeneous variances: the variability of observations near the detection limit is often much greater than the variability of observations well above it. *Gilliom et al.* [1984] explicitly accounted for the increase in variability near the limit of detection in their study of tests for trend.

CONCLUSIONS

The best behaved nonparametric test for comparing median pollution concentrations between two areas based on multiply censored lognormal data is the normal scores test with censored observation scores based on (10) and a permutation variance estimator. This test maintains the nominal α level across a wide range of differing sample sizes and censoring mechanisms.

Several other competing nonparametric tests for comparing median concentrations are available. These tests and the normal scores tests should yield similar results in cases where sample sizes and censoring mechanisms do not differ greatly between areas. In such cases, the Peto-Prentice test with an asymptotic variance estimator may be a more powerful test.

Several software packages already exist to test for median differences based on multiply right-censored data (e.g., BMDP and SAS). These packages can be used to analyze left-censored ground water quality data by multiplying the data by -1 and changing the sign of the resulting test statistics.

REFERENCES

Breslow, N. E., A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship, *Biometrika*, *57*, 579–594, 1970.

Buhler, W. J., The treatment of ties in the Wilcoxon test, *Ann. Math. Stat.*, *38*, 519–523, 1967.

Cohen, A. C., Jr., Simplified estimators for the normal distribution when samples are singly censored or truncated, *Technometrics*, *1*, 217–237, 1959.

Cox, D. R., Regression models and life tables (with discussion), *J. R. Stat. Soc. London, Ser. B*, *34*, 187–220, 1972.

Cox, D. R., and D. V. Hinkley, *Theoretical Statistics*, 511 pp., Chapman and Hall, London, 1974.

Crowley, J., and M. Hu, Covariance analysis of heart transplant data, *J. Am. Stat. Assoc., Assoc.*, *72*, 27–36, 1977.

Deverel, S. J., and S. P. Millard, Distribution and mobility of selenium and other trace elements in shallow ground water of the western San Joaquin Valley, California, *Environ. Sci. Technol.*, *22*, 697–702, 1988.

Deverel, S. J., R. J. Gilliom, R. Fujii, J. A. Izbicki, and J. C. Fields, Areal distribution of selenium and other inorganic constituents in shallow ground water of the San Luis Drain service area, San Joaquin Valley, California: A preliminary study, *U.S. Geol. Surv. Water Resour. Invest. Rep.*, *84-4319*, 1984.

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

Gilbert, R. O., and R. R. Kinnison, Statistical methods for estimating the mean and variance from radionuclide data sets containing negative, unreported or less-than values, *Health Phys.*, *40*, 377–390, 1981.

Gilliom, R. J., and D. R. Helsel, Estimation of distributional parameters for censored trace level water quality data, 1, Estimation techniques, *Water Resour. Res.*, *22*, 135–146, 1986.

Gilliom, R. J., R. M. Hirsch, and E. J. Gilroy, Effect of censoring trace-level water-quality data on trend-detection capability, *Environ. Sci. Technol.*, *18*, 530–535, 1984.

Gleit, A., Estimation for small normal data sets with detection limits, *Environ. Sci. Technol.*, *19*, 1201–1206, 1985.

Helsel, D. R., and R. J. Gilliom, Estimation of distributional parameters for censored trace level water quality data, 2, Verification and applications, *Water Resour. Res.*, *22*, 147–155, 1986.

Hettmansperger, T. P., *Statistical Inference Based on Ranks*, 323 pp., John Wiley, New York, 1984.

Hipel, K. W., Nonparametric approaches to environmental impact assessment, *Water Resour. Bull.*, *24*, 487–492, 1988.

Hirsch, R. M., J. R. Slack, and R. A. Smith, Techniques of trend analysis for monthly water quality data, *Water Resour. Res.*, *18*, 107–121, 1982.

Hollander, M., and D. A. Wolfe, *Nonparametric Statistical Methods*, 503 pp., John Wiley, New York, 1973.

Hughes, J., and S. P. Millard, A tau-like test for trend in the presence of multiple censoring, *Water Resour. Bull.*, *24*, 521–532, 1988.

Kalbfleisch, J. D., and R. L. Prentice, *The Statistical Analysis of Failure Time Data*, 321 pp., John Wiley, New York, 1980.

Kushner, E. J., On determining the statistical parameters for pollution

concentration from a truncated data set, *Atmos. Environ.*, *10*, 975–979, 1976.

Latta, R. B., A monte carlo study of some two-sample rank tests with censored data, *J. Am. Stat. Assoc.*, *76*, 713–719, 1981.

Lee, E. T., *Statistical Methods for Survival Data Analysis*, 557 pp., Lifetime Learning, Belmont, Calif., 1980.

Lehmann, E. L., *Nonparametrics: Statistical Methods Based on Ranks*, 457 pp., Holden-Day, San Francisco, Calif., 1975.

Mantel, N., Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemother. Rep.*, *50*, 163–170, 1966.

Marshall, E., Selenium poisons refuge, California politics, *Science*, *229*, 144–146, 1985.

Miller, R. G., *Survival Analysis*, 238 pp., John Wiley, New York, 1981.

Nehls, G. J., and G. G. Akland, Procedures for handling aromatic data, *J. Air Pollut. Control Assoc.*, *23*, 180–184, 1973.

Owen, W. J., and T. A. DeRouen, Estimation of the mean for lognormal data containing zeroes and left-censored values, with applications to the measurement of worker exposure to air contaminants, *Biometrics*, *36*, 707–719, 1980.

Peto, R., and J. Peto, Asymptotically efficient rank invariant test procedures (with discussion), *J. R. Stat. Soc. London, Ser. A*, *135*, 185–206, 1972.

Pratt, J. W., and J. D. Gibbons, *Concepts of Nonparametric Theory*, 462 pp., Springer-Verlag, New York, 1981.

Prentice, R. L., Linear rank tests with right censored data, *Biometrika*, *65*, 167–179, 1978.

Prentice, R. L., Linear rank tests, in *Encyclopedia of Statistical Science*, vol. 5, pp. 51–58, edited by S. Kotz, and N. L. Johnson, John Wiley, New York, 1985.

Prentice, R. L., and P. Marek, A qualitative discrepancy between censored data rank tests, *Biometrics*, *35*, 861–867, 1979.

Putter, J., The treatment of ties in some nonparametric tests, *Ann. Math. Stat.*, *26*, 368–386, 1955.

S. J. Deverel, Water Resources Division, U.S. Geological Survey, 2800 Cottage Way, Sacramento, CA 95825.

S. P. Millard, CH2M Hill, P.O. Box 91500, Bellevue, WA 98009.

(Received December 28, 1987;
revised July 28, 1988;
accepted August 3, 1988.)