Environ. Sci. Technol. **1998**, *32,* 177−183

Evaluation of Censored Data Methods To Allow Statistical Comparisons among Very Small Samples with Below Detection Limit Observations

JOAN U. CLARKE*

U.S. Army Engineer Waterways Experiment Station, 3909 Halls Ferry Road, Vicksburg, Mississippi 39180

A simulation and verification study assessing the performance of 10 censored data reconstitution methods was conducted to develop guidance for statistical comparisons among very small samples ($n < 10$) with below detection limit observations in dredged sediment testing. Censored data methods were evaluated for preservation of power and nominal type I error rate in subsequent statistical comparisons. Method performance was influenced by amount of censoring, data transformation, population distribution, and variance characteristics. For nearly all situations examined, substitution of a constant such as one-half the detection limit equaled or outperformed more complicated methods. Regression order statistics and maximum likelihood techniques previously recommended for estimating population parameters from censored environmental samples generally performed poorly in very small-sample statistical hypothesis testing with more than minimal censoring, due to their inability to accurately infer distributional properties and their consequent low power or high type I error rates.

Introduction

Chemical analyses of contaminant residues in environmental samples frequently include concentrations reported only as less than a specified limit of detection (LOD). The resulting left-censored data preclude statistical estimation and hypothesis testing by familiar parametric procedures without prior manipulation of the censored observations, although nonparametric procedures may be used directly for certain types of analyses (*1*−*8*). Censored data are sometimes deleted prior to analysis or may be reconstituted using a variety of methods. Simple methods include substituting a constant for all observations below the LOD or using random or evenly spaced numbers from a uniform distribution between zero and LOD for the censored observations. More complicated techniques include distributional methods such as maximum likelihood estimation (MLE) and so-called robust methods such as linear regression on order statistics (ROS) (*9*).

Previous studies have compared various techniques for parameter estimation of censored environmental data sets where sample size $n \geq 10$ (*10*−*22*); most studies recommended MLE or ROS methods. However, little work has been done regarding the performance of censored data methods with very small sample sizes ($n < 10$) typical of environmental evaluations when multiple contaminants are involved and analytical costs are high, as in dredged material contaminant testing. Furthermore, environmental evaluations often necessitate statistical hypothesis testing rather than parameter estimation. Sediment testing for dredging and disposal, for example, may require comparison of contaminant concentrations bioaccumulated from dredged sediments with those bioaccumulated from a reference sediment (*23*, *24*). Elimination of below LOD observations or substitution of a constant such as half the LOD are often used in these types of comparisons because of their simplicity. Other censored data methods have been proposed for use in statistical comparisons (*2*, *4*−*8*, *25*, *26*), but the relative effects of various methods on hypothesis test error rates are generally unknown.

This study was conducted to develop guidance for dredged sediment evaluations in the common situation when some contaminant concentration data are reported as below a single LOD and sample size is very small. The statistical protocol (*24*) calls for comparison of contaminant data from one or more dredged sediment management units with contaminant data from a reference sediment using the Least Significant Difference (LSD) test. Ten censored data reconstitution methods enabling comparisons of censored samples were assessed using simulated data, and the simulation results were verified using chemical concentration data from dredging projects. Criteria for method evaluation included preservation of power and nominal type I error rate in the LSD comparisons.

Experimental Section

Pseudorandom number generators in the SAS statistical package (*27*) were used to create 674 groups of populations. Population parameters were specified for normal, log-normal, and gamma-probability distributions, which are considered likely distributional forms for chemical concentration data in the environment (*10*, *14*, *17*, *28*, *29*). For each of the 674 groups, one population was generated to represent a reference sediment, and one to three additional populations were created to represent dredged sediment treatment(s) that would be compared with the reference. The mean μ_R of each reference population was set to 1, and standard deviation σ_R was set to 0.1, 0.5, 1, 2, or a random number between 0.1 and 2. This encompassed nearly the entire range of coefficients of variation (CVs) calculated for 530 samples of uncensored sediment or tissue contaminant concentration data from several sediment evaluation projects.

Means for the simulated treatment populations (μ_1, μ_2, μ_3) were set equal to 1 for some comparisons to assess type I error rate. For other comparisons, treatment population means were set to a value greater than 1 to assess power; values were chosen such that power was ≈0.5 for the normal distribution, equal variance, uncensored case using untransformed data. Standard deviations for the simulated treatment populations (σ_1, σ_2, σ_3) were set equal to σ_R (i.e., equal variances among populations), equal to the population means (unequal variances proportional to means), or to a random mixture of values between 0.1 and 2 (unequal, mixed variances).

Simulations were conducted using 500 random samples drawn from each population. Sample sizes were equal with either five or eight replicates or unequal with the reference having either the most replicates (reference $n = 6$, treatment $n = 5$, 4, or 3) or the fewest replicates (reference $n = 3$ or 4, treatment $n = 6$). An LOD was imposed at the 20, 40, 60, 80, or 95th percentile of the reference population. The same

* Telephone: 601-634-2954; fax: 601-634-3120; e-mail address: clarkej@ex1.wes.army.mil.

S0013-936X(97)00521-X Not subject to U.S. Copyright. Publ. 1997 Am. Chem. Soc. Published on Web 01/01/1998

TABLE 1. Methods Used in Simulation and Verification Study

method	substitution procedure
UC	none (uncensored comparisons)
DL[a]	constant substitution using LOD (method = CO when used with rankits)
D2[a]	constant substitution using LOD/2 (method = CO when used with rankits)
ZE[a]	constant substitution using zero (method = CO when used with rankits)
UN[b]	uniform distribution: evenly spaced numbers between 0 and LOD
UR	uniform distribution: random numbers between 0 and LOD
LR[b]	ROS: estimation of values from log-normal distribution
NR[b]	ROS: estimation of values from normal distribution
ML[c]	MLE: estimation of values from log-normal distribution using SAS LIFEREG
MN[c]	MLE: estimation of values from normal distribution using SAS LIFEREG
MW[c]	MLE: estimation of values from Weibull distribution using SAS LIFEREG

[a] When rankit transformation is used, DL, D2 and ZE are equivalent (given that all uncensored observations are >LOD), and the method is designated as CO for substitution of any constant between zero and LOD. [b] Procedures described in ref 14. [c] Procedures generally follow ref 26, with the addition of statements to output *n* quantiles of the predicted distribution and substitute the first *c* quantiles for the *c* censored observations in the ordered sample. These modifications were necessary because direct use of the procedure for hypothesis testing (*26*) was not adaptable to the summation of results required for interpretation of millions of simulations.

LOD was then applied to the treatment population(s), with different percentages of those populations affected depending on their means and standard deviations. Although the population censoring percentile was predetermined, each random sample had a variable number of censored observations ranging from 0 to *n*. This "type I" censoring is common in chemical analytical practice when the LOD is known but the number of observations below the LOD varies from sample to sample.

Ten censored data methods for assignment of a numeric value to each censored observation were applied to each group of samples (Table 1). Methods chosen were amenable to simulations using SAS and were considered reasonably uncomplicated for routine regulatory application with the provision of SAS program statements (*30*). Within a set of simulations, all methods were performed on the same samples; new populations were not generated for each method. Following application of each censored data method, treatment samples were compared with the reference using the LSD test (*31*). One of the most powerful multiple comparison procedures, the LSD test is appropriate when control of pairwise rather than experimentwise type I error rate is desired. As regulatory decisions are generally made for each dredged sediment management unit independently of any other management units included in the comparison with a reference, control of experimentwise error rate or "protection" of the LSD test with a prior analysis of variance is unnecessary and could reduce the power of the comparisons.

Because the LSD test assumes normality and equal variances among treatments, $\log_{10}$ and rankit data transformations were used as well as untransformed data. Rankits (normalized ranks or normal scores) are approximations of expected order statistics for the normal distribution and were calculated using the Blom algorithm in SAS (*32*). Log and rankit transformations can normalize samples from log-normal and nonnormal populations, respectively, and often succeed in equalizing variances as well. Subsequent comparisons are tests for differences in geometric means when log transformation is used or for medians when rankits are used. Transformations were applied to the uncensored data and to the reconstituted samples following censoring and use of the censored data methods. One method (LR) reconstituted censored values using logs; these were employed as is for log-transformed data comparisons, back-transformed for untransformed data comparisons, or converted to rankits.

A total of 5 055 000 simulations was performed for each method as well as for uncensored comparisons (UC), and the results for each of the 674 sets of parameters were summarized in terms of actual type I error rate and power. All tests were one-tailed with nominal type I error rate (α) = 0.05. Power was defined as the proportion of statistically significant test results for each method out of 500 simulations given treatment population mean(s) > μ_R. When a treatment population mean = μ_R, the actual type I error rate was the proportion of statistically significant test results for each method, adjusted for the proportion of 500 simulations in which the method could be used. ROS techniques were limited to samples with at least three uncensored observations; MLE techniques could not be used on samples that were completely censored. Other methods were applicable regardless of the amount of censoring.

Upon completion of the simulations, the 10 censored data methods were investigated in a verification study using chemical concentration data from 1079 uncensored sediment and tissue samples analyzed for several dredged material evaluation projects. One to five dredging project samples were simultaneously compared with a reference sample in a total of 786 comparisons involving both equal and unequal *n*. Most sample sizes ranged from three to six replicates, although a few comparisons included samples of up to 12 replicates. The data were artificially censored at the 20, 40, 60, 80, and 95th percentiles of the overall distribution for the samples included in a comparison. Censored data methods were then applied, and LSD comparisons were performed using untransformed, log-transformed, and rankit-transformed data.

Because population parameters of the chemical concentration data sets used for verification were unknown, censored data method results could only be evaluated relative to LSD test results for UC. Power of UC was defined as the number of significant LSD results divided by the total number of comparisons for a given subset of verifications. As the actual type I error rate for UC could not be determined, all significant UC comparisons were assumed to have treatment population mean(s) greater than the reference population mean, and the nominal α was defined to be zero. Power for each censored data method was defined as the fraction of total significant LSD results for UC that was also detected as significant in comparisons following censoring and application of the method, for a given subset of the verification comparisons. Type I error rate for a method was determined as the summation of significant results that were not detected as significant using UC, adjusted for the number of comparisons in which the method could be used. Normality and equality of variances of the uncensored verification data sets were assessed using Shapiro–Wilk's test and Levene's test, respectively, following the protocol for dredged sediment evaluations (*24*). Based on the outcome of these tests,

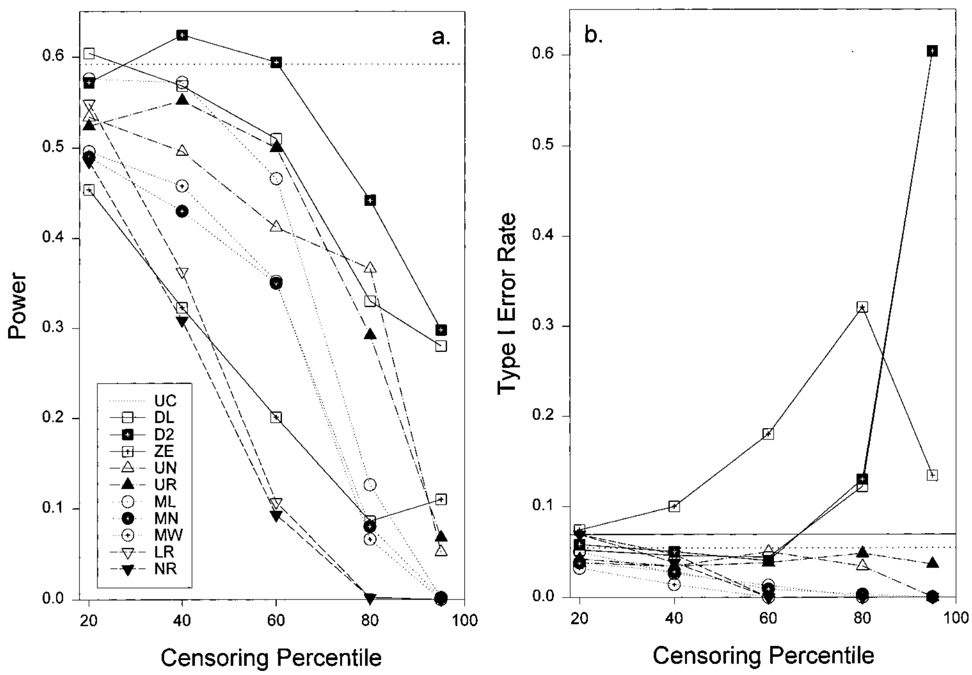

FIGURE 1. Censored data method performance in LSD simulations comparing treatment 1 with reference, using log-transformed samples of $n = 5$ from log-normal distributions where $\sigma_R = \sigma_1 = 0.5$. (a) Power of one-tailed comparison where $\mu_R = 1$ and $\mu_1 = 1.6$. (b) Actual type I error rate of one-tailed comparison where $\mu_R = \mu_1 = 1$; solid line is 95% upper confidence limit (= 0.0691) for nominal α of 0.05.

verification results were examined for subsets of normal, log-normal, and nonnormal distributions and for variances that were equal, proportional to the means, or mixed.

For both the simulation and verification studies, a censored data method was considered unacceptable if it resulted in power less than half that of UC in the LSD test. To avoid comparisons of power among competing methods with dissimilar type I error rates (*33*), methods were eliminated from further consideration if their resulting type I error rates exceeded the upper 95% confidence limit for nominal α calculated using eq 18 in ref 8.

Results

Censored data method performance was influenced strongly by the amount of censoring, as shown in the example of Figure 1. At 20% censoring, power following application of censored data methods was often more similar to that of UC with untransformed data or rankits than with log transformation. Power of the ROS methods declined rapidly as censoring increased, whereas the power of the best methods (usually constant substitution) sometimes exceeded the power of UC, even with censoring as high as 60% (Figure 1a). Beyond 80% censoring, power of most methods was less than half that of UC. Actual type I error rates generally remained acceptable until censoring reached 80−95% (Figure 1b). Note that ZE followed by log transformation causes censored data elimination with attendant loss of power and increase in type I error rate with increased censoring. Censored data elimination has been discouraged due to resulting high bias in statistical estimation procedures and deletion of different proportions of the groups being compared in hypothesis testing procedures (*6*). Thus ZE with log transformation will not be considered further in this paper.

The general performance of the four types of censored data methods (constant substitution, uniform distribution, MLE, and ROS) is summarized in Figure 2. One or more constant substitution method(s) resulted in highest power with acceptable type I error rate in more than 60% of all simulations when censoring did not exceed 80%, but performed acceptably in only one-fifth of all simulations at 95% censoring. Several methods, including uniform distribution and MLE methods, were often tied for highest power at 20−40% censoring (thus, the “best method” bar segments at 20 or 40% censoring in Figure 2 will total >100%). Diminishing performance of the uniform distribution and MLE methods beyond 40% censoring was due primarily to low power rather than to high type I error rates. The ROS methods had either low power or high type I error rates in the majority of simulations at all amounts of censoring. Both power and type I error rates of the ROS methods decreased as censoring increased; power was unacceptably low in all simulations when censoring was 80% or more. The declining power of the MLE and ROS methods as censoring increased was due in large part to the minimum number of uncensored observations per sample required by these methods, and thus their increasing failure rate as the number of trials with unusable samples increased.

Performance of the individual methods was influenced by data transformation, population variance characteristics, and type of distribution in addition to amount of censoring. These factors affected power and type I error rates of the UC as well as the censored data methods. In general, log transformation tended to increase both power and type I error rates as compared with untransformed data and rankits. Unequal variances often resulted in increased type I error rates. The influence of population distribution on individual method performance was inconsistent and difficult to generalize, except that methods tended to perform in a similar manner in the simulations from log-normal and gamma-populations. Within the ranges evaluated in the simulation study, sample size and number of treatments had little effect upon the relative performance of the censored data methods or upon the type I error rates of the UC.

To illustrate a diversity of results while highlighting some previously described general trends, power is presented for treatment 3−reference comparisons having various population variance characteristics, using rankit-transformed samples from gamma-populations (Table 2). Note that UC had relatively high power when $\sigma_R > \sigma_3$ and low power when $\sigma_R < \sigma_3$. MLE methods performed better than constant sub-

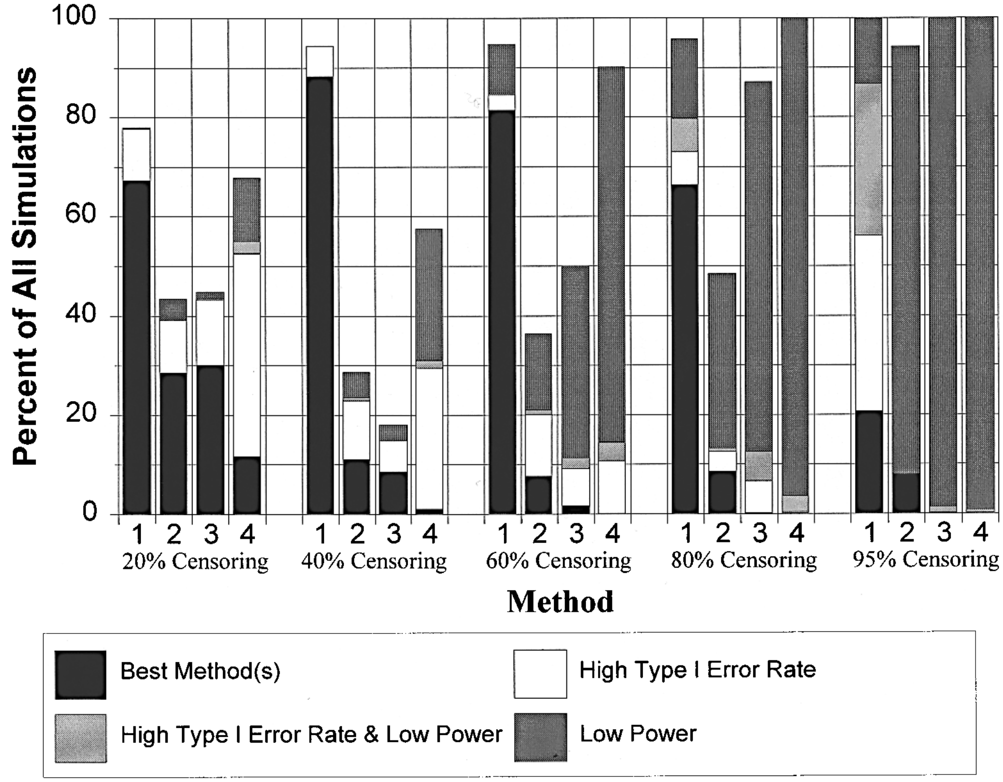

FIGURE 2. General performance of censored data methods in LSD simulations. 1, constant substitution methods; 2, uniform distribution methods; 3, MLE methods; 4, ROS methods. Best methods have highest power with acceptable type I error rate; unacceptable methods have high type I error rate (>0.0691), low power (<half that of UC), or both.

stitution methods in only a few situations of low to moderate censoring with unequal population variances, usually when $\sigma_R \geq 2\sigma_3$. ROS methods outperformed other methods only in a few cases at 20% censoring. One or both uniform distribution methods generally performed well with rankits, often resulting in power similar to CO. Although the best-performing method varied with the simulation conditions, constant substitution methods resulted in highest power in the majority of situations, as expected from Figure 2.

Assumption violations, particularly unequal variances, can profoundly affect both power and type I error rate of parametric statistical comparison procedures (*34, 35*), especially when a control or reference to which other treatments are compared has the largest variance (*30*). Large coefficients of variation influenced LSD test performance even when population variances were equal. Type I error rates for UC exceeded the upper 95% confidence limit for α in 14% of all simulations. These were often cases of mixed variances, especially where the reference coefficient of variation (CV_R) was ≥ 1. UC type I error rates were also high with log-transformed samples from normal populations when variances were mixed, regardless of the magnitude of the CV, or when variances were equal with $CV \geq 1$. High type I error rates for UC were usually translated to the censored data methods at 20% censoring; larger amounts of censoring sometimes lowered the type I error rates of some censored data methods to acceptable levels (Table 2). When variances were proportional to means, type I error rates generally remained low, while the power of the best-performing censored data methods exceeded that of UC and sometimes even increased as censoring increased.

Verifications. Verification results are presented in Table 3 for comparisons grouped by distribution and variance characteristics, using the data transformation that would most likely satisfy the assumptions of the LSD test for each group. The censored data methods resulting in highest power with acceptable type I error rates were D2 with log transformation; D2 or UR with untransformed data; and CO, UN, or UR with rankits. MLE and ROS methods resulted in lower power than the constant substitution and uniform distribution methods at all levels of censoring. In general, no censored data method was satisfactory when censoring exceeded 60%.

Discussion

One or more constant substitution methods performed as well as or better than all other methods in the majority of the simulations and verifications. In most simulations, D2 or DL followed by log transformation, ZE or D2 used with untransformed data, and CO with rankits resulted in higher power than all other methods regardless of amount of censoring. The actual best-performing method in any given case was influenced primarily by the amount of censoring, data transformation, and characteristics of the population variances and distributions, and to a much lesser extent by the limited range of sample sizes and number of treatments included in the study.

Properties of the parent populations are difficult to assess when samples are small and censored. Therefore, the value of recommending different methods for specific data characteristics is limited, and there is justification for continuing the routine practice of substituting a constant such as D2 for very small sample statistical comparisons. Table 4 gives the average power loss, based on the simulations, that would result from using D2 rather than the best-performing censored data method that could be chosen knowing the specific characteristics of the parent populations. In most cases, either D2 was the best method or average power loss was less than 5%. The most substantial power losses occurred with log-transformed data when variances were proportional to means. DL was usually the most powerful method in this situation in the simulations, although the verifications supported the use of D2 (Table 3). DL also performed better

TABLE 2. Power of LSD Treatment 3—Reference Comparisons Using Uncensored Data and Following Application of Censored Data Methods[a]

variances	equal			proportional to means	mixed					
σ_R	0.1	0.5	1.0	1.0	0.4	0.8	1.0	1.0	1.2	1.4
σ_1	0.1	0.5	1.0	1.0	1.8	0.9	0.1	1.4	0.7	1.9
σ_2	0.1	0.5	1.0	1.6	1.9	1.4	2.0	0.1	1.4	0.5
σ_3	0.1	0.5	1.0	2.2	1.3	0.7	0.5	0.2	1.0	1.7
UC	0.528	0.580	0.684[b]	0.260	0.290	0.674	0.730[b]	0.850	0.696[b]	0.474
20% censoring	UN 0.518 CO 0.516 UR 0.516 ML 0.512 MN 0.512 MW 0.512 LR 0.498 NR 0.496	CO 0.594 ML 0.592 MW 0.592 MN 0.586 UN 0.584 UR 0.584	NR 0.684	LR 0.282 UR 0.280 CO 0.270 UN 0.270 MW 0.262 ML 0.260 MN 0.240 NR 0.236	LR 0.466 NR 0.396 CO 0.370 UR 0.350 UN 0.340 MW 0.282 ML 0.278 MN 0.256	CO 0.682 UN 0.678 MW 0.674 ML 0.672 UR 0.668 MN 0.664 LR 0.652 NR 0.632	*c*	MW 0.850 ML 0.848 MN 0.846	*c*	CO 0.500 UN 0.496 MW 0.484 UR 0.482 ML 0.480 MN 0.472 NR 0.462
40% censoring	CO 0.534 UN 0.516 UR 0.502 ML 0.482 MW 0.478 MN 0.476 LR 0.314 NR 0.310	CO 0.586 UN 0.572 UR 0.570 ML 0.558 MW 0.558 MN 0.548 LR 0.328 NR 0.326	CO 0.682 UN 0.672 UR 0.668 MN 0.654 MW 0.654 ML 0.652 LR 0.400 NR 0.395	CO 0.278 UR 0.256 UN 0.254 MW 0.226 ML 0.224 LR 0.167 MN 0.164	CO 0.464 UR 0.400 UN 0.398 LR 0.353 NR 0.300 MW 0.276 ML 0.274 MN 0.242	CO 0.704 UR 0.692 UN 0.678 MW 0.660 ML 0.658 MN 0.638 LR 0.417 NR 0.401	MN 0.732 LR 0.470 NR 0.456	ML 0.844 MN 0.842 NR 0.598	MN 0.654 LR 0.397 NR 0.383	CO 0.498 UR 0.486 UN 0.484 ML 0.470 MW 0.462 MN 0.450 LR 0.274 NR 0.258
60% censoring	CO 0.514 UN 0.470 UR 0.452 ML 0.308 MW 0.308 MN 0.306	CO 0.562 UR 0.520 UN 0.516 ML 0.428 MW 0.412 MN 0.404	CO 0.660 UN 0.626 UR 0.606 ML 0.550 MW 0.532 MN 0.516	CO 0.304 UR 0.256 UN 0.246 ML 0.146 MW 0.144	CO 0.544 UR 0.454 UN 0.448 MW 0.222 ML 0.218 MN 0.182	CO 0.732 UR 0.698 UN 0.682 ML 0.558 MW 0.554 MN 0.530	UN 0.730 UR 0.730 MW 0.672 ML 0.668 MN 0.662	MW 0.802 ML 0.794 MN 0.794	CO 0.658 UN 0.634 UR 0.632 ML 0.556 MN 0.546 MW 0.542	CO 0.472 UN 0.448 ML 0.340 MW 0.334 MN 0.306
80% censoring	CO 0.466 UR 0.354 UN 0.318	CO 0.446 UR 0.336	CO 0.554 UR 0.438 UN 0.390	CO 0.324 UR 0.246 UN 0.166	CO 0.582 UR 0.464 UN 0.442	CO 0.658 UN 0.544 UR 0.536	CO 0.698 UR 0.542 UN 0.530	UN 0.834 MW 0.550 ML 0.548 MN 0.546	CO 0.480 UR 0.390	CO 0.338 UR 0.268
95% censoring	CO 0.370	CO 0.290	*c*	UR 0.166	CO 0.512 UN 0.278	*c*	*c*	*c*	*c*	*c*

[a] Methods listed have power at least half that of UC, and actual type I error rate ≤ 0.0691. Actual type I error rate determined from treatment 1—reference comparison. Rankit-transformed data from gamma-populations. Population means: $\mu_R = 1$, $\mu_1 = 1$, $\mu_2 = 1.6$, $\mu_3 = 2.2$ ($\mu_2 = 1.06$, $\mu_3 = 1.12$ for equal $\sigma = 0.1$; $\mu_2 = 1.3$, $\mu_3 = 1.6$ for equal $\sigma = 0.5$). Equal sample sizes, $n = 5$. [b] Actual type I error rate for UC > 0.0691. [c] All methods have unacceptably high type I error rate or low power.

than other methods in the simulations at low amounts of censoring when variances were equal and CV_R was very low (0.1). With untransformed data, ZE generally performed slightly better than D2 in the simulations but not in the verifications. Constant substitution was the clear choice for use with rankits, although uniform distribution methods often resulted in power similar to CO. However, CO with rankits should generally be preferred to the uniform methods because the latter assign unequal values to censored observations and thus purport to give information that in fact is unknown and possibly incorrect.

Nonparametric procedures such as the two-sample Wilcoxon Rank-Sum test have been recommended for statistical comparisons of censored samples (*6, 8*). Nonparametric rank tests do not require a reconstitution scheme but simply assign tied ranks to all censored observations and are distribution-free, i.e., insensitive to the form of the parent distribution. The use of ranks in parametric procedures was proposed as a bridge between parametric and nonparametric tests (*36*). Rank transformation has the advantage of stabilizing variances; further transformation to rankits imposes normality and can improve test performance to equal that of parametric counterparts (*33*). Universal use of CO with rankits is appealing for small-sample comparisons involving censored data because distributional characteristics are rarely known. In the simulation study, power resulting from CO with rankits averaged 5% lower than that of the best-performing method that could be chosen if log-transformed or untransformed data were known to satisfy the LSD test assumptions. In nine out of the 17 verification combinations with at least one acceptable method (Table 3), CO with rankits resulted in higher power than the best-performing method for the same groups of samples using log-transformed or untransformed data. In the other eight combinations, average power loss for CO with rankits was 6% compared with the best method regardless of transformation. Although the Wilcoxon Rank-Sum test was not included in the simulation study, it should perform similarly to CO with rankits for two-group comparisons (*36*). Both types of tests accurately reflect what is known of the censored data set, in that the censored observations are all somewhere below the LOD but their order is unknown.

MLE and ROS techniques have been recommended for statistical estimation with censored data because of the bias inherent in parameter estimates when constants are substituted for censored observations (*20*). It would seem that censored data methods producing the most accurate estimates of population parameters such as mean and standard deviation should also perform the best when those parameters are used for comparing treatments in statistical hypothesis testing. However, that clearly is not the case when MLE and ROS methods are used to reconstitute censored data for hypothesis testing with very small samples. The low power or high type I error rates resulting from these methods

TABLE 3. Power of LSD Dredged Sediment—Reference Sediment Comparisons Using Uncensored Data and Following Application of Censored Data Methods in Verification Study[a]

distribution	normal			log-normal		nonnormal
variances	equal	proportional to means	mixed	equal or proportional to means	mixed	equal or unequal
transformation	none	log	rankit	log	rankit	rankit
no. of comparisons	268	78	130	128	62	120
95% upper confidence limit for α	0.0261	0.0484	0.0375	0.0378	0.0543	0.0390
UC	0.280	0.449	0.300	0.570	0.387	0.375
20% censoring	D2 0.269 UR 0.261 ZE 0.261 UN 0.250 DL 0.246 ML 0.224 MN 0.220 MW 0.220	D2 0.423 UR 0.423 DL 0.397	CO 0.292 UN 0.292 MN 0.162	D2 0.547 DL 0.531 UR 0.508 UN 0.492 ML 0.328 MW 0.289	CO 0.387 UN 0.387 UR 0.371 ML 0.226 MN 0.226 MW 0.226	CO 0.375 UN 0.375 UR 0.375 ML 0.250 MW 0.250 MN 0.242
40% censoring	UR 0.231 DL 0.220 UN 0.220 ML 0.164 MN 0.160 MW 0.146	D2 0.333	CO 0.246 UN 0.246 UR 0.238	D2 0.477 UR 0.453 DL 0.430 UN 0.422	CO 0.323 UR 0.306 UN 0.290	CO 0.333 UR 0.325 UN 0.317
60% censoring	D2 0.194 ZE 0.187 UR 0.157 UN 0.153	*b*	*b*	D2 0.414 UR 0.383 UN 0.336 DL 0.320	CO 0.258 UR 0.258 UN 0.226	UR 0.217 UN 0.208
80% censoring	*b*	*b*	*b*	*b*	UR 0.194	*b*
95% censoring	*b*	*b*	*b*	*b*	*b*	*b*

[a] Methods listed have power at least half that of UC, and actual type I error rate ≤ 95% upper confidence limit for α. [b] All methods have unacceptably high type I error rate or low power.

TABLE 4. Median Percent Power Loss in LSD Simulations Using D2 Rather than Best-Performing Censored Data Method

transformation		log			none			rankit		
distribution		normal	log-normal	gamma	normal	log-normal	gamma	normal	log-normal	gamma
variances	% censoring									
equal, $CV_R = 0.1$	20	12.4	10.1	5.9	5.7	0.9	1.9	0	0	0
	40	2.6	0.3	0	0	0	0	0	0	0
	60	0	2.8	0	0.5	0.4	0	0	0	0
	80	0	0	0	0	0	0	0	0	0
equal, $CV_R = 0.5$	20	3.3	0	0	1.7	3.3	2.5	0.1	0	0
	40	0	0	0	1.8	5.3	3.4	0	0	0
	60	5.3	0	0	2.0	3.4	3.6	0	0	0
	80	0	0	0	0	1.5	0.4	0	0	0
equal, $CV_R \geq 1.0$	20	2.6	0	0.2	1.4	3.8	1.7	0.5	0	0
	40	5.7	0	0	2.3	5.5	4.6	0	0	0
	60	3.1	0	0	1.4	7.9	6.2	0	0	0
	80	2.0	0	0	0.3	5.7	4.0	0	0	0
proportional to means	20	49.5	4.2	11.4	3.3	1.1	1.4	3.9	0	3.2
	40	28.3	3.5	12.4	7.0	2.1	2.7	0	0	0
	60	19.1	1.7	8.2	6.4	2.3	2.0	0	0	0
	80	10.3	1.6	8.4	4.1	2.9	3.3	0	0	0
mixed	20	13.0	5.1	6.1	0.9	1.4	0.7	0.7	0	0
	40	7.0	0	3.5	1.1	3.5	6.3	0	0	0
	60	0	0	0	1.1	8.9	6.2	0	0	0
	80	0	0	0	4.9	8.2	10.2	0	0	0

make them generally unacceptable when sample size is very small, especially when censoring exceeds 20%.

No censored data method provides a universal panacea for the problem of below detection limit observations in statistical comparisons. All methods result in declining power as censoring increases, eventually reaching a point (power less than half that of UC) where a higher probability of correct decision relative to UC could be obtained by flipping a coin. Simulation results generally place that point around 80% censoring for the best-performing methods, while verification results indicate that none of the methods considered can perform acceptably when censoring exceeds 60%.

The simulation study described herein was designed for single LOD samples. Multiple LODs are common, e.g., when the method LOD is adjusted for the amount of sample matrix available. The effects of censored data methods on hypothesis test power or type I error rate were not assessed when comparisons include multiple LODs, and the results for single LOD simulations cannot be inferred to extend to multiple LOD situations. When constant substitutions such as DL or D2 are used with multiple LODs, the reconstituted data can become functions of sample mass, changing LODs over time, or other factors unrelated to chemical concentration.

Slymen et al. (*26*) recommended a maximum likelihood regression model ("tobit analysis") using the SAS LIFEREG procedure for direct comparison of censored samples without necessitating reconstitution of the unknown observations. Tobit analysis can incorporate probabilities below multiple detection limits and thus should be more appropriate than constant substitution methods for multiple LOD situations. However, like other maximum likelihood censored data methods, tobit analysis may be unsuitable for very small samples due to the lack of information for accurately determining distributional characteristics. Tobit analysis was not evaluated in this study, and its empirical power and type I error rates remain unknown for very small censored sample comparisons.

Acknowledgments

This work was funded by the Dredging Operations Technical Support program of the U.S. Army Corps of Engineers. Permission to publish was granted by the Office, Chief of Engineers. Helpful recommendations on study design were provided by Michael Newman of the Savannah River Ecology Laboratory. The manuscript benefited greatly from insights of three anonymous reviewers and numerous colleagues, especially Richard Kasul, David Moore, and Victor McFarland.

Literature Cited

(1) Gilliom, R. J.; Hirsch, R. M.; Gilroy, E. J. *Environ. Sci. Technol.* **1984**, *18*, 530−535.
(2) Porter, P. S. Ph.D. Thesis, Colorado State University, 1986, 197 pp.
(3) Schneider, H.; Weissfeld, L. *Biometrika* **1986**, *73*, 741−745.
(4) Self, S. G.; Grossman, E. A. *Biometrics* **1986**, *42*, 521−530.
(5) O'Brien, P. C.; Fleming, T. R. *Biometrics* **1987**, *43*, 169−180.
(6) Helsel, D. R. *Environ. Sci. Technol.* **1990**, *24*, 1766−1774.
(7) Atkinson, G. F.; Mount, K. *Can. J. Stat.* **1994**, *22*, 149−162.
(8) Millard, S. P.; Deverel, S. J. *Water Resour. Res.* **1988**, *24*, 2087−2098.
(9) Helsel, D. R.; Hirsch, R. M. *Statistical Methods in Water Resources*; Elsevier: New York, 1992.
(10) Kushner, E. J. *Atmos. Environ.* **1976**, *10*, 975−979.
(11) El-Shaarawi, A. H. *Water Resour. Res.* **1989**, *25*, 685−690.
(12) Gilbert, R. O.; Kinnison, R. R. *Health Phys.* **1981**, *40*, 377−390.
(13) Gleit, A. *Environ. Sci. Technol.* **1985**, *19*, 1201−1206.
(14) Gilliom, R. J.; Helsel, D. R. *Water Resour. Res.* **1986**, *22*, 135−146.
(15) Helsel, D. R.; Gilliom, R. J. *Water Resour. Res.* **1986**, *22*, 147−155.
(16) Helsel, D. R.; Cohn, T. A. *Water Resour. Res.* **1988**, *24*, 1997−2004.
(17) Newman, M. C.; Dixon, P. M.; Looney, B. B.; Pinder, J. E., III. *Water Resour. Bull.* **1989**, *25*, 905−916.
(18) Gaskin, J. E.; Dafoe, T.; Brooksbank, P. *Analyst* **1990**, *115*, 507−510.
(19) Haas, C. N.; Scheff, P. A. *Environ. Sci. Technol.* **1990**, *24*, 912−919.
(20) El-Shaarawi, A. H.; Esterby, S. R. *Water Res.* **1992**, *26*, 835−844.
(21) Hinton, S. W. *Environ. Sci. Technol.* **1993**, *27*, 2247−2249.
(22) Wen, X.-H. *Math. Geol.* **1994**, *26*, 717−731.
(23) U.S. Environmental Protection Agency/U.S. Army Corps of Engineers. *Evaluation of Dredged Material Proposed for Ocean Disposal (Testing Manual)*; EPA-503/8-91/001; Environmental Protection Agency, Office of Marine and Estuarine Protection and Department of the Army, U.S. Army Corps of Engineers: Washington, DC, 1991.
(24) U.S. Environmental Protection Agency/U.S. Army Corps of Engineers. *Evaluation of Dredged Material Proposed for Discharge in Waters of the U. S.*−*Testing Manual (Draft)*; U.S. EPA Office of Water: Washington, DC, 1995.
(25) Prentice, R. L.; Marek, P. *Biometrics* **1979**, *35*, 861−867.
(26) Slymen, D. J.; de Peyster, A.; Donohoe, R. R. *Environ. Sci. Technol.* **1994**, *28*, 898−902.
(27) SAS Institute Inc. *SAS Language Guide, Release 6.03 Edition*; SAS Institute Inc.: Cary, NC, 1988; pp 88−92.
(28) Ott, W. R.; Mage, D. T. *Comput. Ops. Res.* **1976**, *3*, 209−216.
(29) Van Buren, M. A.; Watt, W. E.; Marsalek, J. *Water Res.* **1997**, *31*, 95−104.
(30) Clarke, J. U.; Brandon, D. L. *Applications Guide for Statistical Analyses in Dredged Sediment Evaluations*; Miscellaneous Paper: U.S. Army Engineer Waterways Experiment Station: Vicksburg, MS, in press.
(31) SAS Institute Inc. *SAS/STAT User's Guide, Release 6.03 Edition*; SAS Institute Inc.: Cary, NC, 1988.
(32) SAS Institute Inc. *SAS Procedures Guide, Release 6.03 Edition*; SAS Institute Inc.: Cary, NC, 1988.
(33) Bradley, J. V. *Distribution-free Statistical Tests*; Prentice-Hall, Inc.: Englewood Cliffs, NJ, 1968.
(34) Bradley, J. V. *Br. J. Math. stat. Psychol.* **1978**, *31*, 144−152.
(35) Day, R. W.; Quinn, G. P. *Ecol. Monogr.* **1989**, *59*, 433−463.
(36) Conover, W. J.; Iman, R. L. *Am. Stat.* **1981**, *35*, 124−129.

Received for review June 1, 1997. Revised manuscript received October 8, 1997. Accepted October 14, 1997.[⊗]

ES970521V

[⊗] Abstract published in *Advance ACS Abstracts,* December 1, 1997.