Environ. Sci. Technol. **2007**, *41,* 221−228

Estimating Correlation with Multiply Censored Data Arising from the Adjustment of Singly Censored Data

ELIZABETH NEWTON* AND RUTHANN RUDEL

Silent Spring Institute, 29 Crafts Street, Newton, Massachusetts 02458

Environmental data frequently are left censored due to detection limits of laboratory assay procedures. Left censored means that some of the observations are known only to fall below a censoring point (detection limit). This presents difficulties in statistical analysis of the data. In this paper, we examine methods for estimating the correlation between variables each of which is censored at multiple points. Multiple censoring frequently arises due to adjustment of singly censored laboratory results for physical sample size. We discuss maximum likelihood (ML) estimation of the correlation and introduce a new method (cp.mle2) that, instead of using the multiply censored data directly, relies on ML estimates of the covariance of the singly censored laboratory data. We compare the ML methods with Kendall's tau-b (ck.taub) which is a modification Kendall's tau adjusted for ties, and several commonly used simple substitution methods: correlations estimated with nondetects set to the detection limit divided by 2 and correlations based on detects only (cs.det) with nondetects set to missing. The methods are compared based on simulations and real data. In the simulations, censoring levels are varied from 0 to 90%, ρ from −0.8 to 0.8, and ν (variance of physical sample size) is set to 0 and 0.5, for a total of 550 parameter combinations with 1000 replications at each combination. We find that with increasing levels of censoring most of the correlation methods are highly biased. The simple substitution methods in general tend toward zero if singly censored and one if multiply censored. ck.taub tends toward zero. Least biased is cp.mle2, however, it has higher variance than some of the other estimators. Overall, cs.det performs the worst and cp.mle2 the best.

Introduction

Environmental data frequently are left censored due to detection limits of laboratory assay procedures. Left censored means that some of the observations are known only to fall below a censoring point (detection limit) (*1*). This presents difficulties in statistical analysis of the data (*2*−*4*). Recent papers have discussed correlation estimation when only one of the variables is censored (*5*−*7*), and some attention has been given to the situation when both variables are censored (*8*, *9*). In addition, most work has focused on data that are singly censored (meaning there is only one censoring point). This paper examines methods for estimating the correlation between variables each of which is multiply censored, meaning that there are multiple censoring points.

Multiple censoring can arise if the laboratory detection limit varies from day to day or batch to batch. It can also result from the conversion of a laboratory result to a concentration. For instance, if the laboratory result is micrograms of a particular analyte and this is converted to a concentration by dividing by the physical sample size, for instance grams of dust, then the result will be multiply censored even if the original measurement was not.

Three measures of correlation commonly are used. The Pearson correlation coefficient, r_p, measures the strength of linear association between two variables, x and y. It is equal to the covariance of x and y divided by the product of the standard deviations of x and y.

$$r_p = \frac{\mathrm{Cov}(x,y)}{\{\mathrm{Var}(x)\ \mathrm{Var}(y)\}^{0.5}} \quad (1)$$

Inference about r_p, for small samples, is dependent on the assumption of normality of the data (*10*). When these assumptions are not met, nonparametric methods may be used. These measure the extent of monotone association between two variables.

Two nonparametric methods of measuring correlation are Spearman's rank correlation coefficient, r_s, and Kendall's correlation coefficient, tau. Spearman's r_s is simply the Pearson correlation of the ranks of the data. Kendall's tau is a measure of the concordance of x and y (*11*). Like Spearman's r_s, it is a correlation method based on ranks.

When the data are censored, a common practice is to set the nondetects to the detection limit (DL) divided by a constant, c, (frequently $c = 2$ or the square root of 2) and estimate correlation by standard methods. Another approach has been to set censored values to missing so that correlation is estimated using only simultaneously detected values. This complete-case analysis is known to result in loss of accuracy and precision when the data are not missing completely at random (*12*).

A modification of Kendall's tau for estimating the Kendall correlation in the case of censored data has been suggested by several authors (*4*, *13*, *14*). Here, this method will be referred to as Kendall's tau-b. This is Kendall's tau adjusted for ties, with comparisons involving censored values considered ties under certain conditions. For more information see Helsel (*4*). An example is given in the Supporting Information.

Maximum likelihood (ML) has been advanced by several authors (*6*−*8*) as a method of estimating correlation as well as mean and variance when data are censored. The ML estimate (MLE) of a parameter vector θ is the value of θ that maximizes the likelihood function. In estimating the mean and standard deviation for a variable x, assumed normally distributed, the likelihood function for $\theta = (\mu_x, \sigma_x)$ is

$$L(\theta) = \prod_{i \in I_c} F(Lx_i) \prod_{i \in I_d} f(x_i) \quad (2)$$

where $F(Lx_i)$ is the normal cumulative distribution function (CDF) with parameter θ, $f(x_i)$ is the normal probability density function (PDF) with parameter θ, I_c denotes the set of indices of the censored observations, $x_i < Lx_i$, and I_d denotes the set of indices of the detected observations.

In simultaneously estimating the means, standard deviations and correlation of two variables, x and y, assumed normally distributed, the likelihood function for $\theta = (\mu_x, \mu_y,$

* Corresponding author: phone: 617-332-4288, fax: 617-332-4284, email: newton@silentspring.org.

10.1021/es0608444 CCC: $37.00
Published on Web 11/07/2006

σ_x, σ_y, ρ) is as follows. This equation has been adapted from Lyles et al. (*8*) for multiply censored data.

$$L(\theta) = \prod_{i=1}^{n} G_i \tag{3}$$

where G_i
$= f(x_i, y_i)$ if x_i and y_i are both detected;
$= f(x_i)\ F(Ly_i\ |x_i)$ if x_i is detected and $y_i < Ly_i$;
$= f(y_i)\ F(Lx_i\ |y_i)$ if y_i is detected and $x_i < Lx_i$;
$= F(Lx_i, Ly_i)$ if $x_i < Lx_i$ and $y_i < Ly_i$;
and
Lx_i = detection limit for x_i;
Ly_i = detection limit for y_i;
$f(x_i)$ = normal PDF with parameter (μ_x, σ_x);
$f(y_i)$ = normal PDF with parameter (μ_y, σ_y);
$F(Ly_i\ |x_i)$ = normal CDF with parameter $(\mu_{y|x_i}, \sigma_{y|x})$;
$F(Lx_i\ |y_i)$ = normal CDF with parameter $(\mu_{x|y_i}, \sigma_{x|y})$;
$f(x_i, y_i)$ = bivariate normal PDF with parameter θ;
$F(Lx_i, Ly_i)$ = bivariate normal CDF with parameter θ;
$\mu_{y|x_i} = \mu_y + (\rho\sigma_y/\sigma_x)(x_i - \mu_x)$;
$\mu_{x|y_i} = \mu_x + (\rho\sigma_x/\sigma_y)(y_i - \mu_y)$;
$\sigma_{y|x} = \sigma_y(1 - \rho^2)^{1/2}$;
$\sigma_{x|y} = \sigma_x(1 - \rho^2)^{1/2}$.

Maximizing the likelihood function with respect to all five parameters can be problematic. Convergence may be slow or may fail to occur or may reach a local rather than a global maximum. In a preliminary set of investigations, comparable or superior performance was found when means and standard deviations were estimated separately from the correlation. Examples are shown in the Supporting Information (Figures S41 and S42).

In this investigation, the performance of two ML estimators, denoted cp.mle and cp.mle2, is examined. With cp.mle, the mean and standard deviation of each of two multiply censored variables are estimated using eq 2. Then ρ is estimated using eq 3, with means and standard deviations held fixed. As discussed below, we found this estimator to be biased for multiply censored data.

In situations where the laboratory results (mass of analyte) is singly censored and physical sample size is fully detected, we propose an alternative estimator, cp.mle2. This estimator relies on the following identity (*10*):

$$\begin{aligned}\mathrm{cor}(x - z, y - z) &= \frac{\mathrm{cov}(x,y) - \mathrm{cov}(x,z) - \mathrm{cov}(y,z) + \mathrm{var}(z)}{\{(\mathrm{var}(x) + \mathrm{var}(z) - 2\mathrm{cov}(x,z))(\mathrm{var}(y) + \mathrm{var}(z) - 2\mathrm{cov}(y,z))\}^{0.5}} \\ &= \frac{\sigma_{xy} - \sigma_{xz} - \sigma_{yz} + \sigma_z^2}{\{(\sigma_x^2 + \sigma_z^2 - 2\sigma_{xz})(\sigma_y^2 + \sigma_z^2 - 2\sigma_{yz})\}^{0.5}}\end{aligned} \tag{4}$$

where x = log of laboratory results for one analyte, y = log of laboratory results for another analyte, and z = log of physical sample size. Here, x, y, and z are assumed normally distributed.

With cp.mle2, ML estimates of σ_x, σ_y, σ_z are found separately for each variable using eq 2. Estimates of ρ_{xy}, ρ_{xz}, and ρ_{yz} are found using eq 3. Then $\sigma_{xy} = \rho_{xy}\sigma_x\sigma_y$, $\sigma_{xz} = \rho_{xz}\sigma_x\sigma_z$, and $\sigma_{yz} = \rho_{yz}\sigma_y\sigma_z$.

We show results for simulated data and also for data from the Cape Cod Household Exposure Study. In the Cape Cod Household Exposure Study (CCHES), air, dust, and urine samples were collected from 120 study participants on Cape Cod, Massachusetts (*15*). Urine samples were analyzed for 21 pesticide and phthalate metabolites by the United States Centers for Disease Control using the methods described in ref *16*. For the urine samples, the laboratory reported concentrations of the analytes and, in general, these were singly censored. However to adjust for dilution of the urine, these values were divided by the concentration of creatinine, resulting in multiply censored data. Our efforts to use these data to better understand major sources and pathways of chemical exposure led us to examine methods for estimating correlation between censored variables.

Materials and Methods

In order to compare the performance of correlation estimators, we conducted a simulation experiment and also examined their performance using variables in the CCHES data.

Simulations. Notation: Here x, y, and z (with or without subscripts) are vectors of length n. ρ, ν, p, and q are scalars. $x-z$ is taken element-wise.

The following simulation procedure was repeated 1000 times for each set of parameter values. Data are assumed log normally distributed. Logs of laboratory data (for instance, micrograms of two different analytes, denoted x_{lab} and y_{lab}) are simulated as multivariate normal with mean 0, variance 1, and ρ varying from −0.8 to 0.8. Logs of physical sample sizes (e.g., air volume, dust weight, urine creatinine) denoted z, are simulated as normal, independent of x_{lab} and y_{lab}, with mean 0 and variance, ν, set to either 0 or 0.5. (In the CCHES data, the correlations of laboratory data with physical sample sizes range approximately from −0.1 to 0.6. In this set of simulations, the correlation is assumed to be zero).

Logs of adjusted data (simulated concentrations), then, are $x_{adj} = x_{lab} - z$ and $y_{adj} = y_{lab} - z$. The true Pearson, Spearman, and Kendall correlations between x_{adj} and y_{adj} are computed. From eq 4, the theoretical Pearson correlation of x_{adj} and y_{adj}, $\rho_{adj} = (\rho + \nu)/(1 + \nu)$. The true Spearman correlation is close to this value and the true Kendall correlation is generally 60−80% of ρ_{adj}.

Next, x_{lab} and y_{lab} are censored. Censored proportions of x_{lab}, p_x, and y_{lab}, p_y, are varied from 0.0 to 0.9. For x_{lab}, the sample quantile, q_x, corresponding to p_x is found and regarded as the detection limit. Values of x_{lab} which are less than q_x are set equal to q_x and the singly censored result is denoted x_{sc}. y_{lab} is censored according to the same procedure as x_{lab}. Then the final multiply censored values are $x_{mc} = x_{sc} - z$ and $y_{mc} = y_{sc} - z$.

Seven correlation estimates are computed at each iteration. These are (a) the Pearson correlation with nondetects set to DL/2 (cp.dl2), (b) the Pearson correlation estimated by maximum likelihood using the multiply censored data (cp.mle), (c) the Pearson correlation estimated by maximum likelihood using eq 4 (cp.mle2), (d) the Spearman correlation with nondetects set to DL/2 (cs.dl2), (e) the Spearman correlation based on detects only with nondetects set to missing (cs.det), (f) the Kendall correlation with nondetects set to DL/2 (ck.dl2), (g) Kendall's tau-b (ck.taub). Six of these methods are commonly used and cp.mle2 is proposed as an improved maximum likelihood estimator when unadjusted laboratory data are singly censored.

In the primary set of simulations, the value of the sample size, n, was set to 100. Values of ρ, (the correlation of the x_{lab} and y_{lab}) were varied from −0.8 to 0.8 in increments of 0.4. Values of the variance of the physical sample sizes (denoted ν) were 0 and 0.5. Values of p_x were varied from 0.0 to 0.9, in increments of 0.1. Values of p_y were varied between p_x and 0.9 in increments of 0.1. Thus, the performance of the seven correlation estimates was examined under $5 \times 2 \times 55 = 550$ parameter combinations. There were 1000 replications at each combination. Comparisons were made with sample sizes of 20, 50, and 1000, for a reduced set of parameter combinations.

Simulations and data analysis were carried out using S-Plus version 7.0.4 for Windows (*17*) and R version 2.2.0 for Windows (*18*). The computer programs are available on request.

CCHES Data. We calculated the seven correlation estimates for all 210 pairs of 21 phthalate and pesticide metabolites measured in urine samples of 120 Cape Cod residents. Unlike the simulated data, we do not know the true correlation for these pairs, so our analysis is limited to consideration of (a) consistency among measures, (b) comparison with plotted data and (c) expected correlations based on knowledge about major sources of exposure.

Results and Discussion

Simulations. A good estimator should be both accurate (close to the true value) and precise (have low variability). One robust performance measure that combines these properties is the median absolute deviation (MAD) which here is defined as the median(absolute value(estimate-true value)). (For the Spearman and Kendall correlations, the true value is taken to be the mean of the correlations calculated for uncensored data). Many other performance measures exist and the apparent performance of an estimator will vary depending on the performance measure chosen. Performance of the estimators also varies with the parameter values including the correlation itself. In general, performance is worse with smaller sample sizes and, for multiply censored data, with negative correlations.

Figure 1 shows box plots for each of the estimators with parameter values $n = 100$, $\nu = 0.5$, and $\rho = 0$. The x-axis shows the percent censored in x and y with 55 censoring combinations ranging from (0%,0%) to (90%,90%). The y-axis ranges from -1.0 to 1.0 with the average true correlation indicated. Horizontal lines are drawn at this value $+0.15$ and -0.15. The complete set of box plots for all estimators and parameter combinations, with $n = 100$, is available in the Supporting Information (Figures S1−S30). Additional box plots in the Supporting Information (Figures S31−S40) show the effect of setting the sample size, n, to 20, 50, 100, and 1000 with $\rho = -0.8$, $\nu = 0.5$, for a reduced set of censoring combinations. Here we can see that precision, but in general not accuracy, of the estimators improves with increasing values of n.

In Figure 1, with average Pearson correlation $= (\rho + \nu)/(1 + \nu) = 0.5/1.5 = 0.33$, Spearman correlation $= 0.32$ and Kendall correlation $= 0.22$, we see that the estimators cp.dl2 and cs.det tend away from the true correlation toward more positive values as censoring increases. ck.taub, on the other hand, tends toward zero as censoring increases. cs.det has very high variance with many extreme values. cp.mle2 has the least bias, but higher variance than cp.dl2 and ck.taub.

In general, the behavior of the estimators may be summarized as follows. We emphasize that these results are for simulated data assumed log-normally distributed.

Estimators with Nondetects Set to DL/2. When one of the variables, say x, is censored and the other is not, with increasing levels of censoring, x_{mc} tends toward a constant minus z (log of the physical sample size). The Pearson correlation tends toward $\mathrm{cor}(c - z, y_{sc} - z) = \mathrm{cor}(-z, y_{sc} - z)$, where c is a constant vector. If $\nu = 0$ (the physical sample size, z, is a constant) the correlation tends toward 0. If $\nu = 0.5$, using eq 4 the Pearson correlation tends toward 0.5/sqrt(0.75) = 0.577. The Spearman and Kendall correlations are similar in behavior.

With increasing levels of censoring in both variables, x_{mc} and y_{mc} both tend toward a constant minus z. If $\nu = 0$ these estimators tend toward zero. On the other hand if $\nu \neq 0$, these estimators tend toward one.

cs.det. The behavior of cs.det follows that of cs.dl2, tending toward one if $\nu \neq 0$ and zero if $\nu = 0$. However, the variance of the estimates is much higher. This is the most unreliable of the estimators discussed here and should not be used.

ck.taub. If $\nu = 0$ (data are singly censored) then ck.dl2 and ck.taub give identical results. Even if the data are multiply censored, as the levels of censoring increase, ck.taub, with so many ties in the comparisons, tends toward zero.

Maximum Likelihood Estimators. If $\nu = 0$ (data are singly censored) then cp.mle and cp.mle2 give the same results. However, if the data are multiply censored then cp.mle is negatively biased at high levels of censoring. cp.mle2, on the other hand, has little bias. The variance tends to be higher than that of many of the other methods, however.

The performance of these ML estimators could be improved if better estimates of the mean and standard deviation could be obtained. Preliminary work indicates that imputation methods discussed in (*1*) and Kaplan-Meier methods can achieve greater accuracy and precision than ML in estimating these parameters. This is an avenue for future research.

Median Absolute Deviation (Mad) Results. Table 1 shows the maximum levels of censoring (in both x and y) for each parameter combination investigated which result in MAD < 0.1. Here, we can see that cs.det performs the worst, seldom achieving MAD < 0.1. For all sample sizes and correlation estimators, the worst performance is found for negatively correlated multiply censored data ($\rho = -0.8$, $\nu = 0.5$). Outside of this situation, for samples of size $n = 100$, methods with nondetects set to DL/2 achieve MAD < 0.1 with 40−80 or 90% censoring, Kendall's tau-b with 50−80% censoring and cp.mle2 with 70−90% censoring. For $n = 50$, the Kendall correlations were the most consistent with MAD < 0.1 for 50−70% censoring, cp.mle2 achieved MAD < 0.1 with 20−90% censoring. For samples of size $n = 20$, the estimators achieved MAD < 0.1 only with highly positive correlations. For $\rho = 0.4$, $\nu = 0.5$, only the Kendall correlations achieved MAD < 0.1 for up to 40% censoring. It should be noted that, for the same dataset, Kendall correlations in general are smaller in magnitude than Spearman or Kendall so the criterion of MAD < 0.1 actually favors the Kendall correlations slightly.

Correlation Estimates with CCHES Urinary Metabolites Data. In the CCHES urine data, there are 21 different analytes and $21 \times 20/2 = 210$ pair wise relationships among them. We examined the distribution of each variable using normal probability plots and examined the relationships among them using scatter plots and then compared the performance of the seven estimates of correlation discussed above. Here we discuss four examples. All analyses were conducted using logs of the data.

Figure 2 shows normal probability plots (QQ plots) for seven of the analytes and also for creatinine. Here the data are unadjusted and largely singly censored. Figure 3 shows scatter plots of the creatinine adjusted data for each of the selected pairs. Table 2 shows the numbers detected, censored and missing for each variable and Table 3 shows the seven correlation estimates for each relationship discussed.

The normal probability plots in Figure 2 were created using the S-Plus Environmental Statistics function qqplot-.censored. The logs of the ordered data (empirical quantiles) are plotted on the vertical axis and the corresponding theoretical quantiles from the assumed normal distribution are plotted on the horizontal axis. This is described further in Millard (*1*). Here we can see that the variables appear to satisfy the assumption of a normal distribution with the possible exception of 2Naph and OPP.

In the scatter plots, nondetects are plotted at the detection limit. Each point is represented by a letter. "B" indicates that both are detected, "x" that x is detected, "y" that y is detected and "n" that neither is detected. Hence, for a point represented by an x, if the true value were known, the plotting position would be somewhere on a vertical line extending below the x. For a point represented by a y, if the true value were known, the plotting position would be somewhere on a horizontal line extending to the left of the y. For a point

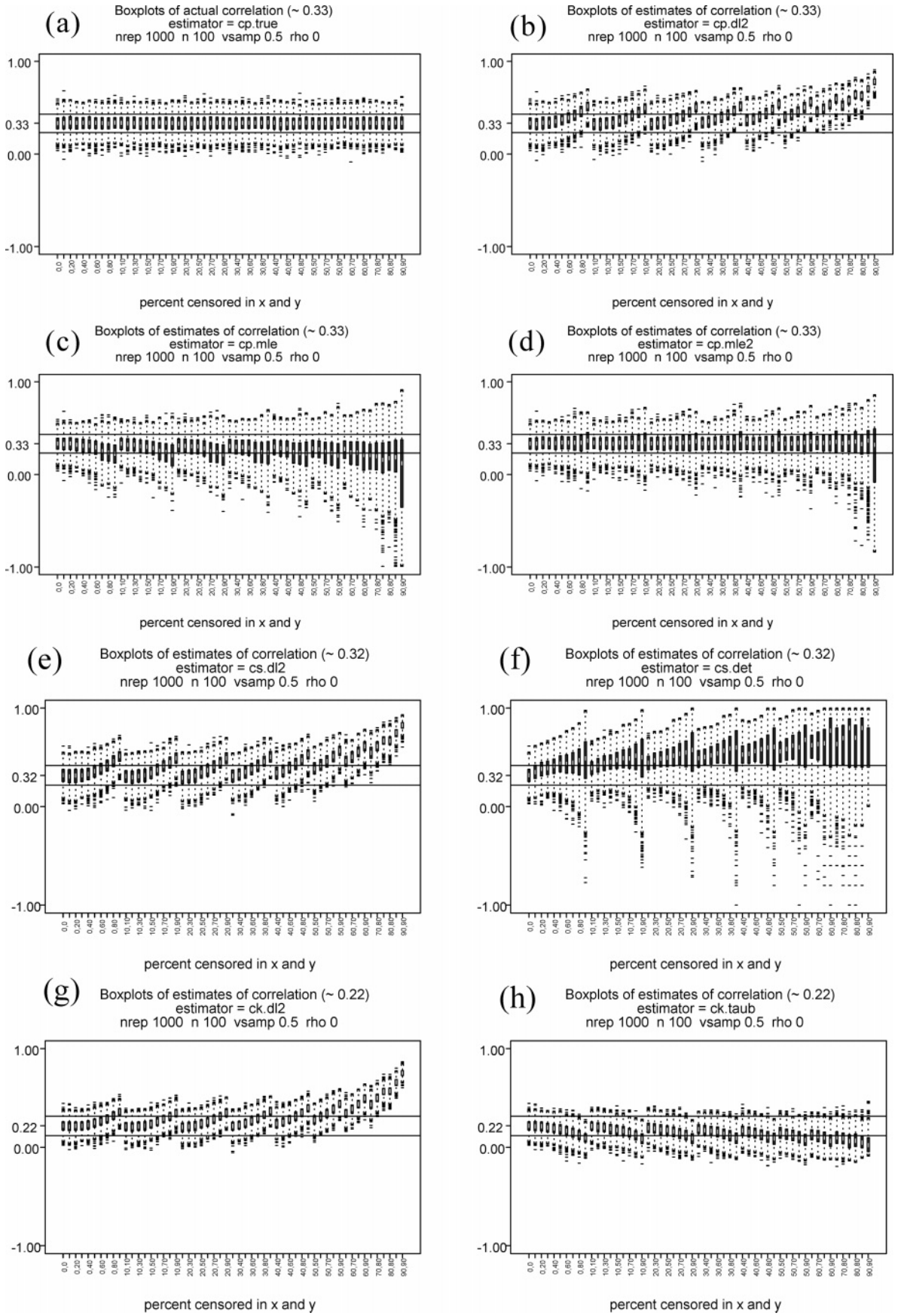

FIGURE 1. Simulation results. Box plots of estimates of correlation for parameter values: $n = 100$, $\rho = 0$, $\nu = 0.5$, number of repetitions = 1000. Horizontal axis shows percent censored in *x* and *y*. (a) Pearson correlation for uncensored data (average is 0.33). (b) Estimator is cp.dl2. (c) Estimator is cp.mle. (d) Estimator is cp.mle2. (e) Estimator is cs.dl2 (average uncensored Spearman correlation is 0.32). (f) Estimator is cs.det. (g) Estimator is ck.dl2 (average uncensored Kendall correlation is 0.22). (h) Estimator is ck.taub.

TABLE 1. Simulation Results. Maximum Levels of Censoring (in both *x* and *y*) for Each Parameter Combination Which Results in Median Absolute Deviation <0.1[a]

	cp.dl2	cp.mle	cp.mle2	cs.dl2	cs.det	ck.dl2	ck.taub
$n = 100$, $\rho = 0.8$, $\nu = 0$	70	90	90	60	10	80	80
$n = 100$, $\rho = 0.4$, $\nu = 0$	60	70	70	60	10	80	80
$n = 100$, $\rho = 0.0$, $\nu = 0$	90	70	70	80	20	80	80
$n = 100$, $\rho = -0.4$, $\nu = 0$	50	60	60	50	0	80	80
$n = 100$, $\rho = -0.8$, $\nu = 0$	30	60	60	40	0	50	50
$n = 100$, $\rho = 0.8$, $\nu = 0.5$	90	90	90	90	80	70	80
$n = 100$, $\rho = 0.4$, $\nu = 0.5$	70	60	90	60	50	60	60
$n = 100$, $\rho = 0.0$, $\nu = 0.5$	50	60	70	50	10	50	60
$n = 100$, $\rho = -0.4$, $\nu = 0.5$	40	30	70	40	0	40	50
$n = 100$, $\rho = -0.8$, $\nu = 0.5$	30	20	60	40	0	40	30
$n = 20$, $\rho = 0.8$, $\nu = 0.5$	90	80	90	90	40	70	60
$n = 20$, $\rho = 0.4$, $\nu = 0.5$	0	0	0	0	0	40	40
$n = 20$, $\rho = 0$, $\nu = 0.5$	0	0	0	0	0	0	0
$n = 20$, $\rho = -0.4$, $\nu = 0.5$	0	0	0	0	0	0	0
$n = 20$, $\rho = -0.8$, $\nu = 0.5$	0	0	0	0	0	0	0
$n = 50$, $\rho = 0.8$, $\nu = 0.5$	90	90	90	90	70	70	70
$n = 50$, $\rho = 0.4$, $\nu = 0.5$	70	60	70	60	30	60	60
$n = 50$, $\rho = 0$, $\nu = 0.5$	40	20	30	40	0	50	50
$n = 50$, $\rho = -0.4$, $\nu = 0.5$	30	0	20	10	0	40	50
$n = 50$, $\rho = -0.8$, $\nu = 0.5$	30	10	30	30	0	40	30
$n = 1000$, $\rho = -0.8$, $\nu = 0.5$	30	20	70*	40	0	40	30

[a] 70% censoring in both x and y was the maximum censoring level tested with samples of size $n = 1000$.

represented by an n, if the true value were known, the plotting position would be somewhere below and to the left of the n.

Figure 3a shows the relationship between monoethyl phthalate (MEP) and monobutyl phthalate (MBuP), which are urinary monoester metabolites of diethyl phthalate and di-n-butyl phthalate. We expect these two phthalates to be positively correlated because the parent compounds are used in personal care products such as fragrances and cosmetics (*19*). With one missing and no censored values in MEP and only two censored values in MBuP there are 117 simultaneous detects. As might be expected, the correlation estimates are consistent (Pearson estimates 0.25, cs.dl2 0.23, Kendall estimates 0.16). cs.det (0.21) is lower than cs.dl2 because it does not take into account the influence of the two points in the lower left of the plot that are censored for MBuP.

Figure 3b shows the relationship between 2,4-dichlorophenol (24DCP) and 2,5-dichlorophenol (25DCP). We expect these urinary metabolites to be positively correlated because they are both derived from exposure to chlorinated benzenes and chlorinated phenols. 94 values (78%) are censored in 24DCP and 21 values (18%) are censored in 25DCP resulting in only 25 simultaneous detects and 20 simultaneous nondetects. As might be expected the correlation estimates vary widely. Based on the simulation results, we expect that ck.taub (0.25) underestimates the true Kendall correlation at this level of censoring. Here cp.mle2 is 0.46.

Next we discuss two pairs of urinary metabolites that are not expected to be highly correlated based on major sources of exposure. Figure 2c shows a plot of 2-naphthol (2Naph), a metabolite of the polyaromatic hydrocarbon naphthalene, and isopropoxyphenol (IPP), a metabolite of the pesticide propoxur. 2Naph has 78% censored values and IPP has 60% censored values resulting in only 10 simultaneous detects and 56 simultaneous nondetects. The plot shows a diagonal line of n's where neither compound was detected. (Just as a reminder, these are plots of the creatinine-adjusted data and simultaneous nondetects are plotted as "n" at the detection limit divided by the concentration of creatinine). Correlation methods based on setting censored values to the DL/2 thus can be artificially inflated. (This artifact is shown also in plots of simulated data in the Supporting Information, Figure S43). Kendall's tau-b and the ML estimates are not vulnerable to this error and are close to zero. The estimate based on detects only (0.46) is high because of the small number of detects and the strong influence of the single point in the upper right corner.

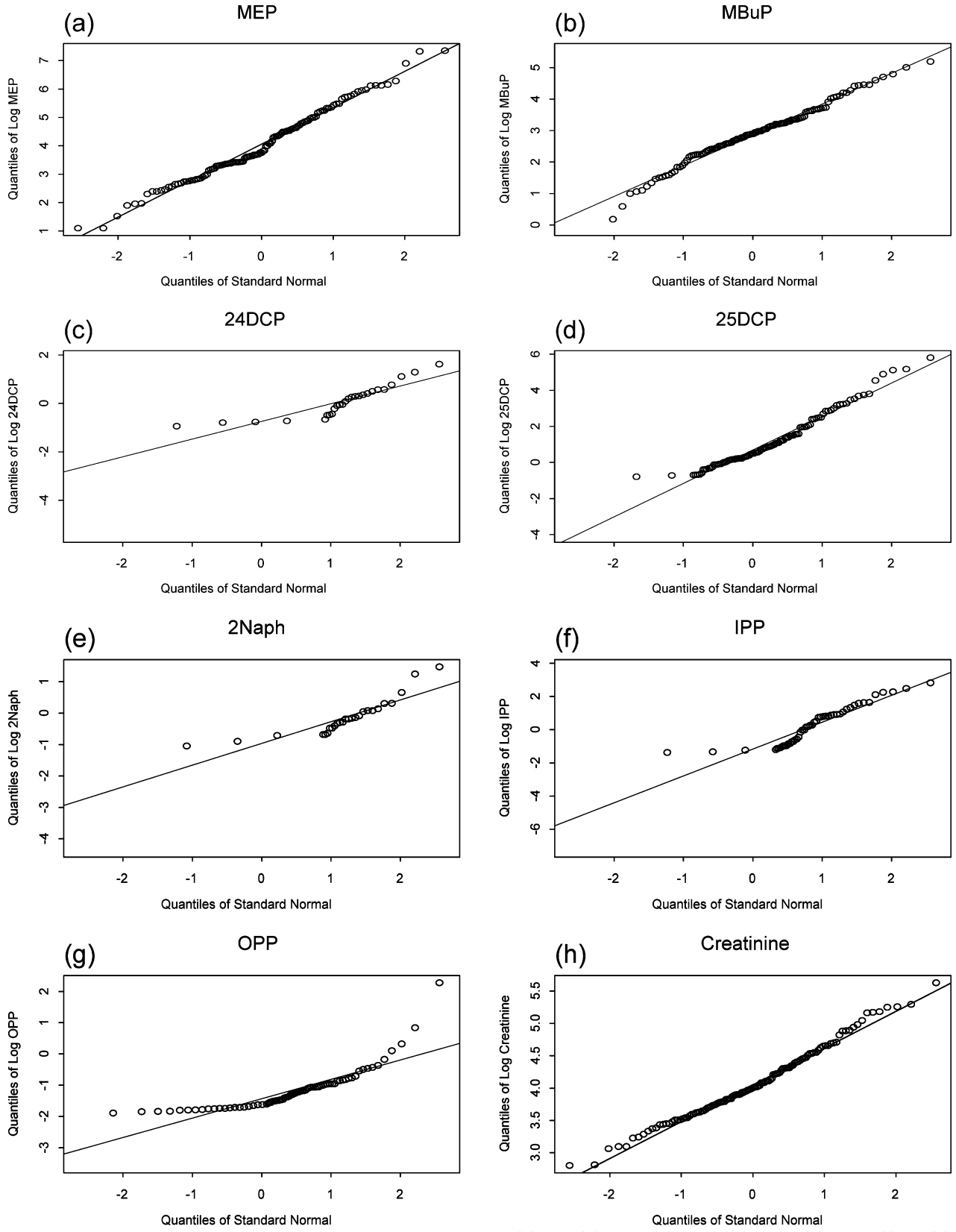

FIGURE 2. Normal probability plots of logs of selected CCHES variables. (a) MEP. (b) MBuP. (c) 24DCP. (d) 25DCP. (e) 2Naph. (f) IPP. (g) OPP. (h) Creatinine.

Figure 2d shows a plot of the pesticide metabolite 2,4-dichlorophenol (24DCP) and the disinfectant o-phenyl phenol (OPP). 24DCP has 78% censored values and OPP has 36% censored values resulting in 20 simultaneous detects and 37 simultaneous nondetects. Again, the correlation estimates based on setting the nondetects to DL/2 and cs.det are high. The ML methods and ck.taub give much lower estimates (0.05 or less).

Summary and Recommendations. Always plot the data. QQ plots investigate the distribution of the data. Scatter plots show relationships between variables and can reveal outliers and influential points.

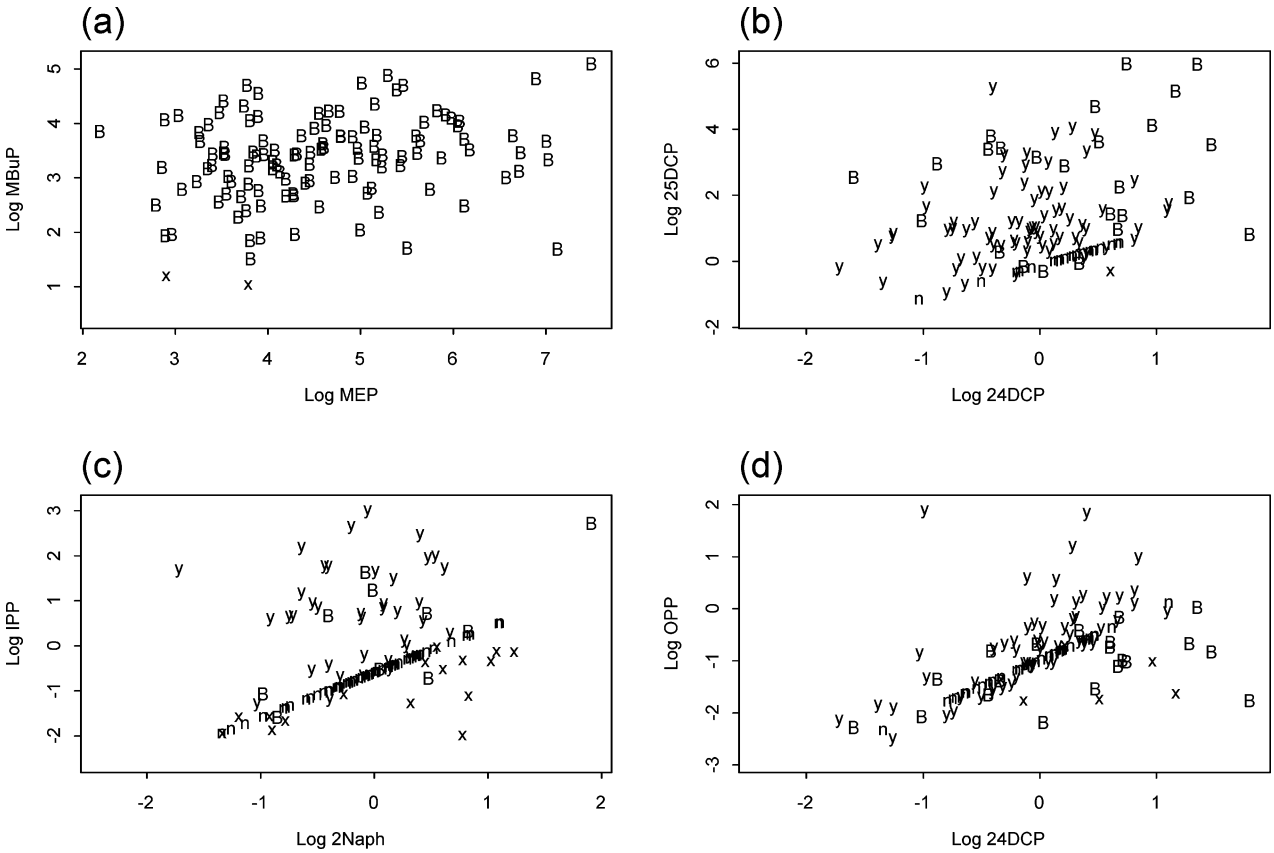

FIGURE 3. Scatter plots of selected CCHES variables, nondetects are plotted at detection limit. For each observation, B indicates both variables are detected, x indicates *x* is detected, y indicates *y* is detected n indicates neither is detected. (a) *x* is log MEP, *y* is log MBuP. (b) *x* is log 24DCP, *y* is log 25DCP. (c) *x* is log 2Naph, *y* is log IPP. (d) *x* is log 24DCP, *y* is log OPP.

TABLE 2. Numbers Detected, Censored, and Missing in Selected CCHES Variables

X	Y	n[a]	ndx[b]	ncx[c]	nmx[d]	pcx[e]	ndy[f]	ncy[g]	nmy[h]	pcy[i]	nsimdet[j]
MEP	MBuP	119	119	0	1	0.0	118	2	0	1.7	117
24DCP	25DCP	120	26	94	0	78.3	99	21	0	17.5	25
2Naph	IPP	120	26	94	0	78.3	48	72	0	60.0	10
24DCP	OPP	120	26	94	0	78.3	77	43	0	35.8	20

[a] Number not missing in x and y. [b] Number detected in x. [c] Number censored in x. [d] Number missing in x. [e] Percent censored in x. [f] Number detected in y. [g] Number censored in y. [h] Number missing in y. [i] Percent censored in y. [j] Number simultaneously detected in x and y.

TABLE 3. Correlation Estimates for Selected CCHES Variables

X	Y	cp.dl2	cp.mle	cp.mle2	cs.dl2	cs.det	ck.dl2	ck.taub
MEP	MBuP	0.25	0.25	0.24	0.23	0.21	0.16	0.16
24DCP	25DCP	0.34	0.01	0.46	0.21	0.24	0.15	0.25
2Naph	IPP	0.17	−0.01	0.00	0.32	0.46	0.28	−0.01
24DCP	OPP	0.32	0.02	0.07	0.45	0.45	0.35	0.05

Correlations using detects only never should be used.

For samples of size 20, the correlation estimators achieve MAD < 0.1, only for highly positive correlations.

For samples of size 50, ck.taub gives the most consistent results, achieving MAD < 0.1 with 50−70% censoring in most cases.

For samples of size 100 or more, cp.mle2 gives the best results and can be used with 60 to 90% censoring, depending, unfortunately, on the correlation.

Because the behavior of the estimators is complicated and depends on the value of the parameter being estimated, we suggest comparing estimates of correlation and not relying too heavily on any single estimate for highly censored data.

Future Work. In future work, we will examine the impact of departures from the assumptions employed in the simulations, in particular, the assumption that the data are log normally distributed. We also would like to look more closely at the effect of sample size.

Here we have focused on point estimates of correlation. Future work will examine interval estimates.

As mentioned above, the performance of maximum likelihood estimators of correlation would be enhanced if better estimates of the mean and variance could be obtained. Future work will investigate parametric and nonparametric (for instance Kaplan−Meier) methods for improving estimates of mean and variance.

Statistical methods must be developed that explicitly incorporate the variability of all observations and not simply that of the censored values. The analysis of duplicate samples can help to assess the variability of the laboratory methods.

Acknowledgments

This research was supported by grants from the Hurricane Voices Breast Cancer Foundation, the Heinz Endowments, the National Cancer Institute (grant no. 5 R03 CA103478-02) and the National Institute of Environmental Health Sciences (grant no. 1 R25 ES013258-01). We thank anonymous reviewers for many helpful suggestions and comments.

Supporting Information Available

Box plots of simulation results for all parameter values considered, including results with different sample sizes and results with simultaneous estimation of all five parameters for cp.mle; scatter plots of simulated data and an example of the computation of Kendall's tau-b. This material is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) Millard, S.; Neerchal, N. *Environmental Statistics with S-Plus*; CRC Press: Boca Raton, FL, 2001.

(2) Helsel, D. R. Less than obvious: Statistical treatment of data below detection limit. *Environ. Sci. Technol.* **1990**, 1767−1774.

(3) Helsel, D. R. More than obvious. Better methods for interpreting nondetect data. *Environ. Sci. Technol.* **2005a**, 419A−423A.

(4) Helsel, D. R. *Nondetects and Data Analysis*; John Wiley and Sons, Inc.: Hoboken, NJ, 2005b.

(5) Lynn, H. Maximum likelihood inference for left-censored HIV RNA data. *Stat. Med.* 2001, *20*, 35−45.

(6) Lyles, R. H.; Fan, D.; Chuachoowong, R. Correlation coefficient estimation involving a left censored laboratory assay variable. *Stat. Med.* **2001a**, *20*, 2921−2933.

(7) Song, J.; Barnhart, H. X.; Lyles, R. H. A GEE approach for estimating correlation coefficients involving left-censored variables. *J. Data Sci.* **2004**, *2*, 245−257.

(8) Lyles, R. H.; Williams, J. K.; Chuachoowong, R. Correlating two viral load assays with known detection limits. *Biometrics* **2001b**, *57*, 1238−1244.

(9) Benning, L.; Lyles, R. H.; Gange, S. J. Methods for comparing correlations involving left-censored laboratory data. In *ASA Proceedings of Joint Statistical Meetings, Section on Statistics in Epidemiology*; Alexandria, VA, 2002; pp. 212−216.

(10) Tamhane, A.; Dunlop, D. *Statistics and Data Analysis, from Elementary to Intermediate*; Prentice Hall, Inc.: Upper Saddle River, NJ, 2000.

(11) Hollander, M.; Wolfe, D. *Nonparametric Statistical Methods,* 2nd ed.; John Wiley and Sons, Inc.: New York, 1999.

(12) Little, R.; Rubin, D. *Statistical Analysis with Missing Data*, 2nd ed.; John Wiley and Sons, Inc.: Hoboken, NJ, 2002

(13) Oakes, D. A concordance test for independence in the presence of censoring. *Biometrics* **1982**, *38*, 451−455.

(14) Brown, B. W.; Hollander, M.; Korwar, R. M. Nonparametric tests of independence for censored data, with applications to heart transplant studies. *Reliab. Biometry* **1974**, 327−354.

(15) Rudel, R. A.; Camann, D. E.; Spengler, J. D.; Korn, L. R.; Brody, J. G. Phthalates, alkylphenols, pesticides, polybrominated diphenyl ethers, and other endocrine disrupting compounds in indoor air and dust. *Environ. Sci. Technol.* **2003**, *37*, 4543−4553.

(16) Centers for Disease Control and Prevention *Third National Report on Human Exposure to Environmental Chemicals*; National Center for Environmental Health Division of Laboratory Science: Atlanta, GA, 2005.

(17) Insightful Corporation; S-PLUS Version 7.0.4 for Microsoft Windows ed., 2005.

(18) The R Foundation for Statistical Computing; R Version 2.2.0 (2005−10.-06 r35749) ed., 2005.

(19) Hauser, R.; Calafat, A. M. Phthalates and human health. *Occup. Environ. Med.* **2005**, *62*, 806−818.

Received for review April 7, 2006. Revised manuscript received September 21, 2006. Accepted September 27, 2006.

ES0608444