Supporting Information

Censoring trace-level environmental data: statistical analysis considerations to limit bias

Barbara Jane George[1*], Leslie Gains-Germain[2], Kristin Broms[2], Kelly Black[2], Marschall Furman[3], Michael D. Hays[4], Kent W. Thomas[1], Jane Ellen Simmons[1]

[1] Center for Public Health and Environmental Assessment, Office of Research and Development, U.S. EPA, Research Triangle Park, North Carolina 27711, United States

[2] Neptune and Company, Inc., Lakewood, Colorado 80215, United States

[3] Oak Ridge Institute for Science and Education (ORISE) Research Participant at U.S. EPA, Office of Research and Development, Center for Public Health and Environmental Assessment, Research Triangle Park, North Carolina 27711, United States

[4] Center for Environmental Measurement and Modeling, Office of Research and Development, U.S. EPA, Research Triangle Park, North Carolina 27711, United States

*Address correspondence to B.J. George, CPHEA/ORD/U.S. EPA, 109 T.W. Alexander Dr., Research Triangle Park, NC 27711 USA. Telephone: (919) 541-4551. E-mail: george.bj@epa.gov

Table of Contents

SW-846 Method Detection Limit, Revision 1, July 1992

This MDL definition that was applied in the original cook-stove study was not cited by Shen et al. 2016 and is provided here because we recommend, as stated in the concluding sentence of the paper, "report and publish: all measurement data without censoring along with data quality indicators, detection limits, reporting levels and other censoring thresholds, and the methods used in calculating these, without regard to the use of censoring in both primary and secondary analyses."

METHOD DETECTION LIMIT (MDL):

The minimum concentration of a substance that can be measured and reported with 99% confidence that the analyte concentration is greater than zero and is determined from analysis of a sample in a given matrix type containing the analyte.

For operational purposes, when it is necessary to determine the MDL in the matrix, the MDL should be determined by multiplying the appropriate one-sided 99% t-statistic by the standard deviation obtained from a minimum of three analyses of a matrix spike containing the analyte of interest at a concentration three to five times the estimated MDL, where the t-statistic is obtained from standard references or the table below.

No. of samples:	t-statistic
3	6.96
4	4.54
5	3.75
6	3.36
7	3.14
8	3.00
9	2.90
10	2.82

Estimate the MDL as follows:
Obtain the concentration value that corresponds to:

a) an instrument signal/noise ratio within the range of 2.5 to 5.0, or

b) the region of the standard curve where there is a significant change in sensitivity (i.e., a break in the slope of the standard curve).

Determine the variance (S^2) for each analyte as follows:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

where x_i = the ith measurement of the variable x and $\bar{x}$ = the average value of x;

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Determine the standard deviation (s) for each analyte as follows:

$$s = (S^2)^{1/2}$$

Determine the MDL for each analyte as follows:

$$MDL = t_{(n-1,\ \alpha = .99)}(s)$$

where $t_{(n-1,\ \alpha = .99)}$ is the one-sided t-statistic appropriate for the number of samples used to determine (s), at the 99 percent level.

Example S1. Use of R *EnvStats:: elnormAltCensored()* for Maximum Likelihood Estimation (MLE) and R *EnvStats::enparCensored()* for Kaplan Meier estimation of the mean and SD of 10 data observations where 1 observation has been censored.

Step 1. Randomly generate 10 lognormal(zeta, sigma) observations using SAS data step code adapted from https://blogs.sas.com/content/iml/2017/05/10/simulate-lognormal-data-sas.html

```
%let N = 10;        /* sample size */
data LN10;
call streaminit(98765);
sigma = 0.5;     /* shape or log-scale parameter    */
zeta = 2;        /* scale or log-location parameter */
do i = 1 to &N;
  Y = rand("Normal", zeta, sigma);  /* Y ~ N(zeta, sigma)    */
  X = exp(Y);                  /* X ~ LogN(zeta, sigma) */
  output;
end;
keep X;
run;
```

Step 2. After censoring the smallest observation (or if starting with previously censored data), prepare the data for the R *EnvStats* functions *enparCensored() and elnormAltCensored().* Here the data generated in Step 1 are in the "orig.X" column. These data are stored in a file named "data10, 1 obs censored.csv."

orig.X	X	censored
<5	5	TRUE
5.5852	5.5852	FALSE
6.2982	6.2982	FALSE
6.7822	6.7822	FALSE
7.4194	7.4194	FALSE
7.4619	7.4619	FALSE
8.9753	8.9753	FALSE
9.7251	9.7251	FALSE
10.1279	10.1279	FALSE
14.6743	14.6743	FALSE

Step 3. Execute R code

```
setwd ("C:\\MLE & KM example\\")    #update the path information before execution
getwd()

library(EnvStats)     #R package EnvStats must be installed prior to execution

data <- read.csv("data10, 1 obs censored.csv")     #update the data's CSV filename, as warranted
summary(data)
```

```
X <- data$X
X

censored <- data$censored
censored

elnormAltCensored(X,censored,method="mle")
# in line above, replace “mle” with “rROS” for Robust Regression on Order Statistics estimation

enparCensored(X,censored)
```

Step 4. Maximum Likelihood Estimation output from Step 3.

```
Results of Distribution Parameter Estimation
Based on Type I Censored Data
--------------------------------------------

Assumed Distribution:            Lognormal

Censoring Side:                  left

Censoring Level(s):              5

Estimated Parameter(s):          mean = 8.1337926
                                 cv   = 0.3399739

Estimation Method:               MLE

Data:                            X

Censoring Variable:              censored

Sample Size:                     10

Percent Censored:                10%
```

Noting that cv = SD/mean, SD = mean*cv = 2.765

Step 5. Kaplan Meier output from Step 3.

```
Results of Distribution Parameter Estimation
Based on Type I Censored Data
--------------------------------------------

Assumed Distribution:            None

Censoring Side:                  left

Censoring Level(s):              5

Estimated Parameter(s):          mean    = 8.2049500
                                 sd      = 2.6910675
                                 se.mean = 0.8509903

Estimation Method:               Kaplan-Meier

Data:                            X

Censoring Variable:              censored

Sample Size:                     10

Percent Censored:                10%
```

Step 6. Here are the true sample mean and SD from the SAS MEANS procedure. Note the censored data value was 3.8816.

Analysis Variable : X

N	Mean	Std Dev
10	8.093112	2.994668

Step 7. Calculate Bias: compare MLE and Kaplan Meier estimates to the mean and SD from the full, uncensored data. Note that censoring removed information that is not truly reconstituted by statistical analysis.

Bias is the estimated value minus the true value

	Descriptive Statistics		Bias (Estimated - True value)	
Approach	Mean	SD	Mean	SD
Full sample (n=10 observations)	8.093[a]	2.995[a]		
After censoring 1 observation				
Maximum Likelihood Estimation	8.134[b]	2.765[b]	0.041	-0.230
Kaplan-Meier	8.205[b]	2.691[b]	0.112	-0.304

[a]True value

[b]Estimated value

Table S1. Method detection limit (MDL) and calibration curve lowest value (CCLV) with sample sizes[a] by category for each PAH considered for the case study

PAH	PAH Abbreviation	MDL (ng/µL)	CCLV (ng/µL)	n < MDL	n between MDL & CCLV	n ≥ CCLV
Benzo(a)pyrene	BaP	0.010	0.05		8	44
Benzo(b)fluoranthene	BbF	0.016	0.10	2	6	44
Benzo(e)pyrene	BeP	0.010	0.05		8	44
Benzo(g,h,i)pyrene	BghiP	0.019	0.05	1	3	46
Benzo(k)fluoranthene	BkF	0.011	0.05	2	4	46
Coronene	COR	0.025	0.05		5	47
Dibenzo[a,h]anthracene	DBA	0.023	0.10	8	18	21
Indeno[1,2,3-c,d]pyrene	IcdP	0.040	0.05	7		44
Perylene	PER	0.011	0.05	1	4	41

[a]Sample size counts omit blanks and samples for which measurement values could not be recorded (i.e., signal-to-noise ratio <=1)

Table S2.GC-MS dibenzo(a,h)anthracene (DBA) data used in the case study (n=47) and corresponding method detection limit (MDL) and calibration curve lowest value (CCLV)

DBA (ng/µL)	MDL (ng/µL)	CCLV (ng/µL)	Detection Flag
0.01	0.023	0.10	< MDL
0.01	0.023	0.10	< MDL
0.01	0.023	0.10	< MDL
0.01	0.023	0.10	< MDL
0.01	0.023	0.10	< MDL
0.01	0.023	0.10	< MDL
0.02	0.023	0.10	< MDL
0.02	0.023	0.10	< MDL
0.03	0.023	0.10	< CCLV
0.03	0.023	0.10	< CCLV
0.04	0.023	0.10	< CCLV
0.04	0.023	0.10	< CCLV
0.04	0.023	0.10	< CCLV
0.04	0.023	0.10	< CCLV
0.05	0.023	0.10	< CCLV
0.05	0.023	0.10	< CCLV
0.05	0.023	0.10	< CCLV
0.06	0.023	0.10	< CCLV
0.07	0.023	0.10	< CCLV
0.07	0.023	0.10	< CCLV
0.07	0.023	0.10	< CCLV
0.07	0.023	0.10	< CCLV
0.07	0.023	0.10	< CCLV
0.07	0.023	0.10	< CCLV
0.08	0.023	0.10	< CCLV
0.08	0.023	0.10	< CCLV
0.10	0.023	0.10	Detect
0.11	0.023	0.10	Detect
0.12	0.023	0.10	Detect
0.12	0.023	0.10	Detect
0.15	0.023	0.10	Detect
0.15	0.023	0.10	Detect
0.19	0.023	0.10	Detect
0.32	0.023	0.10	Detect
0.40	0.023	0.10	Detect
0.52	0.023	0.10	Detect
0.60	0.023	0.10	Detect
0.75	0.023	0.10	Detect

0.87	0.023	0.10	Detect
0.89	0.023	0.10	Detect
1.00	0.023	0.10	Detect
1.12	0.023	0.10	Detect
1.16	0.023	0.10	Detect
1.31	0.023	0.10	Detect
1.31	0.023	0.10	Detect
1.46	0.023	0.10	Detect
2.94	0.023	0.10	Detect

Table S3. R syntax used to estimate mean and standard deviation for approaches used in the case study

Approach	R Syntax	Supports Single/Multiple Censoring
Uncensored		
Full sample	mean(), sd()	Both
Full sample Maximum Likelihood Estimation[a]	elnormAlt(,method="mle")$parameters[b]	Both
After censoring		
Substitute MDL/2	mean(), sd()[c]	Both
Maximum Likelihood Estimation[a]	elnormAltCensored(,censored,method="mle")$parameters[d]	Both
Robust Regression on Order Statistics[a]	elnormAltCensored(,censored,method="rROS")$parameters[e]	Both
Kaplan-Meier[a]	enparCensored()[f]	Both
After omitting observations		
Complete case	mean(), sd()	Both

[a]*EnvStats* package https://cran.r-project.org/web/packages/EnvStats/EnvStats.pdf. This *EnvStats* documentation describes function options, including which of its methods support only singly censored data and which support multiply censored data.

[b]elnormAlt() assumes a lognormal distribution and uses maximum likelihood estimation for the mean and coefficient of variation. The standard deviation is estimated by the product of this mean and coefficient of variation. Maximum likelihood estimation is a parametric method based on maximizing the likelihood function.

[c]sd() function is based on variance estimated using n-1 as the divisor

[d]elnormAltCensored(method="mle") assumes a lognormal distribution and uses maximum likelihood estimation to estimate the mean and coefficient of variation in the presence of censored data. The standard deviation is estimated by the product of the mean and coefficient of variation. Maximum likelihood estimation is a parametric method based on maximizing the likelihood function. Similarly, use egammaAltCensored() for censored data when assuming a gamma distribution.

[e]elnormAltCensored(method="rROS") robust regression on order statistics is considered semiparametric because values are imputed parametrically for the censored data and combined with (observed) detected values, which are not parametric, to complete the distribution. The imputed values are predicted by fitting parametric quantile-quantile regression on the log-transformed data, back-transforming their values to original scale, and estimating the mean and standard deviation using method of moments. For more information, see *EnvStats* documentation

for the elnormAltCensored() function https://cran.r-project.org/web/packages/EnvStats/EnvStats.pdf and page 15-15 in US EPA (2009) Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities, Unified Guidance, *EPA-530-R-09*-007, Office of Resource Conservation and Recovery Program Implementation and Information Division, U.S. Environmental Protection Agency, Washington, https://archive.epa.gov/epawaste/hazard/web/html/index-12.html, accessed July 19, 2020.

[f]enparCensored() Kaplan-Meier estimator is nonparametric because it uses the observed data's empirical cumulative distribution function (ECDF) directly (rather than fitting a parametric distribution). The fitted ECDF is used to estimate the population mean and standard deviation adjusted for data censoring. The enparCensored() default settings, used here, include *censoring.side="left"* and the generalized product-limit probability method for plotting positions due to Michael and Schucany (1986) for estimating the ECDF. Several probability methods for calculating the plotting positions (empirical probabilities) are available and include Michael and Schucany (1986) generalized product-limit (the default), Gillespie et al. (2010) reverse Kaplan Meier, and Hirsch and Stedinger (1987) generalized product-limit. For more information, see *EnvStats* documentation for the enparCensored() and ecdfPlotCensored() functions https://cran.r-project.org/web/packages/EnvStats/EnvStats.pdf and page 15-7 in US EPA (2009) Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities, Unified Guidance, *EPA-530-R-09*-007, Office of Resource Conservation and Recovery Program Implementation and Information Division, U.S. Environmental Protection Agency, Washington, https://archive.epa.gov/epawaste/hazard/web/html/index-12.html, accessed July 19, 2020. Also see:

Michael, J.R., and W.R. Schucany (1986). Analysis of Data from Censored Samples, In D'Agostino, R.B., and M.A. Stephens, eds. Goodness-of Fit Techniques. Marcel Dekker, New York, 560pp, Chapter 11, 461-496.

Gillespie, B.W., Q. Chen, H. Reichert, A. Franzblau, E. Hedgeman, J. Lepkowski, P. Adriaens, A. Demond, W. Luksemburg, and D.H. Garabrant. (2010). Estimating Population Distributions When Some Data Are Below a Limit of Detection by Using a Reverse Kaplan-Meier Estimator. Epidemiology 21(4), S64–S70.

Hirsch, R.M., and J.R. Stedinger. (1987). Plotting Positions for Historical Floods and Their Precision. Water Resources Research 23(4), 715-727.

Table S4. GC-MS dibenzo(a,h)anthracene (DBA) data used in the group comparisons: counts before resampling (n=24), corresponding method detection limit (MDL), calibration curve lowest value (CCLV), and counts after resampling for each distinct DBA value (n=85)

Stove Type	Stove	DBA (ng/µL)	MDL (ng/µL)	CCLV (ng/µL)	Resampled Counts
Stove Type 2 (n=40 after resampling)	Butterfly Model (n=9)	0.01	0.023	0.10	8
		0.02	0.023	0.10	3
		0.06	0.023	0.10	4
		0.07	0.023	0.10	7
		0.08	0.023	0.10	4
		0.10	0.023	0.10	2
		0.11	0.023	0.10	4
		0.12	0.023	0.10	3
		0.19	0.023	0.10	5
Stove Type 1 (n=45 after resampling)	Eco Chula XXL (n=15)	0.01	0.023	0.10	5
		0.01	0.023	0.10	
		0.03	0.023	0.10	2
		0.04	0.023	0.10	9
		0.04	0.023	0.10	
		0.05	0.023	0.10	2
		0.07	0.023	0.10	7
		0.07	0.023	0.10	
		0.07	0.023	0.10	
		0.07	0.023	0.10	
		0.12	0.023	0.10	6
		0.15	0.023	0.10	6
		0.15	0.023	0.10	
		0.32	0.023	0.10	3
		0.52	0.023	0.10	5
	Jiko Poa Rocket[a] (n=19)	0.01	0.023	0.10	
		0.01	0.023	0.10	
		0.02	0.023	0.10	
		0.04	0.023	0.10	
		0.04	0.023	0.10	
		0.07	0.023	0.10	
		0.08	0.023	0.10	
		0.40	0.023	0.10	
		0.60	0.023	0.10	
		0.75	0.023	0.10	
		0.87	0.023	0.10	

		0.89	0.023	0.10
		1.00	0.023	0.10
		1.12	0.023	0.10
		1.16	0.023	0.10
		1.31	0.023	0.10
		1.31	0.023	0.10
		1.46	0.023	0.10
		2.94	0.023	0.10
	Solgas/Repsol[a]	0.01	0.023	0.10
	(n=4)	0.03	0.023	0.10
		0.05	0.023	0.10
		0.05	0.023	0.10

[a]Jiko Poa Rocket and Solgas/Repsol stoves were omitted from the group comparisons.

Table S5. R syntax used to test differences in groups of resampled stove data

Approach	R Syntax
Uncensored	
Full sample	geom_stripchart(test.text=TRUE)[a]
Maximum Likelihood Estimation	cenmle(obs, censored, groups)[b,c]
After censoring at CCLV	
Substitute CCLV/2	geom_stripchart(test.text=TRUE)[a]
Maximum Likelihood Estimation	cenmle(obs, censored, groups)[b]

[a] *EnvStats* package https://cran.r-project.org/web/packages/EnvStats/EnvStats.pdf. The geom_stripchart(test.text=TRUE) function and option specifies one dimensional scatter plots of the data groups; their sample sizes (n), means, 95% confidence intervals for the means, standard deviations (SD); and a two-sample t test for difference of means (i.e., that the difference is zero) assuming normally distributed data.

[b] *NADA* package https://cran.r-project.org/web/packages/NADA/NADA.pdf. The cenmle() assumes lognormally distributed data and the regression coefficient estimate for "groups" is the magnitude of the difference between the two groups expressed as the ratio of their geometric means, with *p*-value from test that the difference is zero. Alternatively cenfit() estimates the group medians and cendiff() nonparametrically tests the difference between groups. For more information see https://www.practicalstats.com/resources/NADA-resources/NADAforR_Examples.pdf, page 17.

[c] Logical value of FALSE assigned to 'censored' for all observations

Table S6. Akaike Information Criterion (AIC) statistics for the case study

Data	n[a]	Normal	Lognormal	Gamma
Full sample[b]	47	83.9	-27.1	-15.3
Censored at MDL[c]	39	144.5	34.8	39.1
Censored at CCLV[c]	21	203.3	93.9	90.2

[a] Count of observations not censored
[b] AIC calculated using fitdist() function in R *fitdistrplus* package
[c] AIC calculated using fitdistcens() function in R *fitdistrplus* package

Table S7. Bias (ng/µL) in the case study mean and SD when censoring at the MDL

Approach	Bias in mean relative to		Bias in SD relative to	
	Full sample	MLE[a]	Full sample	MLE[a]
After censoring 8 observations				
Substitute MDL/2	0.000	n/a[b]	0.000	n/a[b]
Maximum Likelihood Estimation[c]	0.071	0.041	1.105	0.384
Robust Regression on Order Statistics	0.082	0.052	1.202	0.481
Kaplan-Meier	0.002	-0.028	-0.007	-0.728
After omitting 8 observations				
Complete case	0.070	n/a[b]	0.033	n/a[b]

[a]Maximum likelihood estimate for the full sample assuming lognormal distribution

[b]Bias not estimated because approach is inconsistent with lognormal distributional assumption

[c]Maximum likelihood estimation after censoring assuming lognormal distribution

Table S8. Bias (ng/µL) in the case study mean and SD when censoring at the CCLV

Approach	Bias in mean relative to		Bias in SD relative to	
	Full sample	MLE[a]	Full sample	MLE[a]
After censoring 26 observations				
Substitute CCLV/2	0.004	n/a[b]	-0.003	n/a[b]
Maximum Likelihood Estimation[c]	0.182	0.152	2.932	2.211
Maximum Likelihood Estimation[d]	-0.012	-0.042	0.093	-0.628
Robust Regression on Order Statistics	0.046	0.016	0.471	-0.251
Kaplan-Meier	0.032	0.002	-0.023	-0.745
After omitting 26 observations				
Complete case	0.387	n/a[b]	0.112	n/a[b]

[a]Maximum likelihood estimation for the full sample assuming lognormal distribution

[b]Bias not estimated because approach is inconsistent with lognormal distributional assumption

[c]Maximum likelihood estimation after censoring assuming lognormal distribution

[d]Maximum likelihood estimation after censoring assuming gamma distribution

Table S9. Results of distribution fitting using AIC for 1,000 samples from moderately skewed lognormal(1, 0.5)

	Censoring level, n=20			Censoring level, n=50		
	30%	50%	80%	30%	50%	80%
Proportion Lognormal selected						
Full sample[a]	0.63	0.65	0.67	0.72	0.73	0.72
Substitute censoring-level/2[a]	0.39	1.00	1.00	0.46	1.00	1.00
Maximum Likelihood Estimation[b]	0.46	0.37	0.29	0.55	0.50	0.39
Proportion Gamma selected						
Full sample [a]	0.32	0.30	0.28	0.27	0.27	0.27
Substitute censoring level/2 [a]	0.58	0	0	0.54	0	0
Maximum Likelihood Estimation[b]	0.42	0.54	0.56	0.44	0.49	0.58

[a]AIC calculated using fitdist() function in R *fitdistrplus* package

[b]AIC calculated for censored data using fitdistcens() function in R *fitdistrplus* package

Table S10. Results of distribution fitting using AIC for 1,000 samples from highly skewed lognormal(-2.2, 1.6)

	Censoring level, n=20			Censoring level, n=50		
	30%	50%	80%	30%	50%	80%
Proportion Lognormal selected						
Full sample[a]	0.81	0.80	0.80	0.95	0.94	0.94
Substitute censoring-level/2[a]	0.98	1.00	1.00	1.00	1.00	1.00
Maximum Likelihood Estimation	0.61	0.51	0.30	0.81	0.74	0.50
Proportion Gamma selected						
Full sample	0.19	0.20	0.20	0.06	0.06	0.06
Substitute censoring level/2	0.02	0	0	0	0	0
Maximum Likelihood Estimation	0.39	0.49	0.70	0.19	0.27	0.50

[a]AIC calculated using fitdist() function in R *fitdistrplus* package

[b]AIC calculated for censored data using fitdistcens() function in R *fitdistrplus* package

Table S11. Bias in the mean of moderately skewed lognormal(1,0.5) distribution from simulation study

Approach	n	Censoring Level (%)	Number Samples	Average	Min	Max
Full Sample[a]	20	30	1000	-0.002	-0.989	1.443
		50	1000	0.000	-1.116	1.240
		80	1000	0.015	-0.872	1.593
	50	30	1000	0.016	-0.615	1.088
		50	1000	-0.001	-0.766	0.842
		80	1000	0.002	-0.740	0.754
Full Sample Maximum Likelihood Estimation[a]	20	30	1000	0.000	-0.979	1.563
		50	1000	0.003	-1.118	1.115
		80	1000	0.017	-0.853	1.611
	50	30	1000	0.017	-0.598	1.010
		50	1000	0.000	-0.755	0.906
		80	1000	0.004	-0.742	0.710
Substitute CensoringLevel/2[b]	20	30	1000	-0.171	-1.128	1.352
		50	1000	-0.284	-1.367	1.026
		80	1000	-0.289	-1.308	1.575
	50	30	1000	-0.145	-0.760	0.829
		50	1000	-0.278	-1.025	0.595
		80	1000	-0.296	-1.127	0.717
Maximum Likelihood Estimation[b]	20	30	1000	-0.032	-1.033	1.458
		50	1000	-0.042	-1.236	1.436

		80	1000	0.049	-1.372	2.786
	50	30	1000	-0.005	-0.606	0.936
		50	1000	-0.038	-0.904	0.899
		80	1000	-0.029	-1.088	1.744
Robust Regression on Order Statistics[b,c]	20	30	583	-0.041	-1.012	1.276
		50	458	-0.064	-0.993	1.164
		80	441	0.012	-1.454	2.784
	50	30	560	0.012	-0.589	0.899
		50	511	-0.054	-0.727	0.853
		80	424	-0.134	-1.044	1.287
Kaplan-Meier[b]	20	30	1000	0.148	-0.895	1.757
		50	1000	0.400	-0.952	2.113
		80	1000	1.391	-0.267	4.498
	50	30	1000	0.171	-0.470	1.215
		50	1000	0.403	-0.517	1.410
		80	1000	1.374	0.028	3.111
Complete Case[b]	20	30	1000	0.621	-0.664	2.672
		50	1000	1.143	-0.485	3.189
		80	1000	2.475	0.257	6.857
	50	30	1000	0.661	-0.179	1.952
		50	1000	1.160	-0.102	2.644
		80	1000	2.492	0.853	5.119

[a]Full, uncensored sample

[b]After censoring

[c]Bias was estimated only for samples where lognormal was the selected distribution because the R *EnvStats* package does not currently support gamma distributions for this approach

Table S12. Bias in the mean of highly skewed lognormal(-2.2,1.6) distribution from simulation study

Approach	n	Censoring Level (%)	Number Samples	Average	Min	Max
Full Sample[a]	20	30	1000	0.002	-0.316	9.263
		50	1000	-0.004	-0.334	3.842
		80	1000	0.008	-0.333	10.163
	50	30	1000	0.001	-0.273	1.753
		50	1000	-0.001	-0.272	1.754
		80	1000	0.002	-0.275	2.179
Full Sample Maximum Likelihood Estimation[a]	20	30	1000	0.037	-0.316	3.843
		50	1000	0.044	-0.336	2.402
		80	1000	0.041	-0.330	3.427
	50	30	1000	0.010	-0.250	0.750
		50	1000	0.008	-0.274	0.878
		80	1000	0.013	-0.280	1.079
Substitute CensoringLevel/2[b]	20	30	1000	0.002	-0.317	9.262
		50	1000	0.002	-0.334	3.852
		80	1000	0.101	-0.332	10.343
	50	30	1000	0.001	-0.273	1.751
		50	1000	0.005	-0.273	1.761
		80	1000	0.087	-0.257	2.252

Maximum Likelihood Estimation[b]	20	30	1000	0.011	-0.317	4.759
		50	1000	0.004	-0.329	2.292
		80	1000	0.099	-0.338	73.882
	50	30	1000	0.001	-0.273	1.234
		50	1000	-0.002	-0.278	1.357
		80	1000	0.000	-0.279	2.344
Robust Regression on Order Statistics[b,c]	20	30	611	0.058	-0.314	9.260
		50	506	0.048	-0.336	3.822
		80	300	0.078	-0.338	10.095
	50	30	807	0.020	-0.272	1.749
		50	735	0.025	-0.250	1.738
		80	500	0.032	-0.290	2.106
Kaplan-Meier[b]	20	30	1000	0.010	-0.312	9.265
		50	1000	0.034	-0.328	3.891
		80	1000	0.295	-0.294	10.608
	50	30	1000	0.009	-0.266	1.756
		50	1000	0.034	-0.260	1.796
		80	1000	0.265	-0.201	2.422
Complete Case[b]	20	30	1000	0.162	-0.290	13.398
		50	1000	0.340	-0.281	8.023
		80	1000	1.133	-0.252	51.987
	50	30	1000	0.162	-0.228	2.666
		50	1000	0.351	-0.176	3.849
		80	1000	1.138	0.027	12.001

[a]Full, uncensored sample

[b]After censoring

[c]Bias was estimated only for samples where lognormal was the selected distribution because the R *EnvStats* package does not currently support gamma distributions for this approach

Table S13. Bias in the standard deviation of moderately skewed lognormal(1,0.5) distribution from simulation study

Approach	n	Censoring Level (%)	Number Samples	Average	Min	Max
Full Sample[a]	20	30	1000	-0.063	-1.050	2.131
		50	1000	-0.062	-1.018	2.147
		80	1000	-0.038	-0.910	2.494
	50	30	1000	-0.010	-0.747	1.554
		50	1000	-0.030	-0.796	1.557
		80	1000	-0.035	-0.731	1.506
Full Sample Maximum Likelihood Estimation[a]	20	30	1000	-0.056	-1.055	2.076
		50	1000	-0.045	-1.006	1.429
		80	1000	-0.032	-0.915	1.594
	50	30	1000	0.002	-0.639	1.318
		50	1000	-0.017	-0.718	0.975
		80	1000	-0.018	-0.698	0.973
Substitute CensoringLevel/2[b]	20	30	1000	0.104	-0.795	2.215
		50	1000	0.128	-0.779	2.316
		80	1000	-0.112	-0.890	2.373
	50	30	1000	0.143	-0.548	1.658

		50	1000	0.149	-0.613	1.670
		80	1000	-0.099	-0.658	1.443
Maximum Likelihood Estimation[b]	20	30	1000	-0.084	-1.060	1.963
		50	1000	-0.059	-1.082	1.800
		80	1000	-0.080	-1.455	2.423
	50	30	1000	-0.024	-0.834	1.100
		50	1000	-0.023	-0.841	1.198
		80	1000	-0.028	-1.122	1.797
Robust Regression on Order Statistics[b,c]	20	30	583	0.027	-1.076	2.181
		50	458	-0.007	-1.063	2.335
		80	441	-0.171	-1.599	2.828
	50	30	560	0.108	-0.664	1.604
		50	511	0.051	-0.770	1.664
		80	424	0.016	-1.126	1.782
Kaplan-Meier[b]	20	30	1000	-0.253	-1.168	1.948
		50	1000	-0.435	-1.306	1.795
		80	1000	-0.887	-1.589	1.725
	50	30	1000	-0.175	-0.980	1.409
		50	1000	-0.367	-1.122	1.319
		80	1000	-0.819	-1.486	0.856
Complete Case[b]	20	30	1000	-0.181	-1.236	2.540
		50	1000	-0.240	-1.285	2.955
		80	1000	-0.317	-1.623	5.755
	50	30	1000	-0.118	-0.994	1.778

		50	1000	-0.170	-1.113	2.302
		80	1000	-0.242	-1.412	3.709

[a]Full, uncensored sample

[b]After censoring

[c]Bias was estimated only for samples where lognormal was the selected distribution because the R *EnvStats* package does not currently support gamma distributions for this approach

Table S14. Bias in the standard deviation of highly skewed lognormal(-2.2,1.6) distribution from simulation study

Approach	n	Censoring Level (%)	Number Samples	Average	Min	Max
Full Sample[a]	20	30	1000	-0.574	-1.307	40.526
		50	1000	-0.606	-1.290	16.458
		80	1000	-0.538	-1.328	44.131
	50	30	1000	-0.434	-1.205	10.682
		50	1000	-0.426	-1.264	11.025
		80	1000	-0.421	-1.205	13.580
Full Sample Maximum Likelihood Estimation[a]	20	30	1000	0.547	-1.281	105.100
		50	1000	0.588	-1.242	33.690
		80	1000	0.640	-1.309	122.072
	50	30	1000	0.137	-1.137	7.692
		50	1000	0.139	-1.185	9.776
		80	1000	0.174	-1.146	13.974
Substitute CensoringLevel/2[b]	20	30	1000	-0.574	-1.306	40.527

		50	1000	-0.610	-1.294	16.456
		80	1000	-0.591	-1.342	44.087
	50	30	1000	-0.435	-1.204	10.682
		50	1000	-0.429	-1.269	11.024
		80	1000	-0.458	-1.229	13.567
Maximum Likelihood Estimation[b]	20	30	1000	0.724	-1.302	367.790
		50	1000	0.433	-1.296	98.637
		80	1000	268.041	-1.343	264318.781
	50	30	1000	0.099	-1.209	25.532
		50	1000	0.151	-1.257	22.015
		80	1000	0.962	-1.197	221.689
Robust Regression on Order Statistics[b,c]	20	30	611	-0.336	-1.257	40.527
		50	506	-0.337	-1.270	16.463
		80	300	-0.023	-1.344	44.147
	50	30	807	-0.322	-1.199	10.683
		50	735	-0.269	-1.182	11.028
		80	500	-0.111	-1.171	13.592
Kaplan-Meier[b]	20	30	1000	-0.599	-1.312	39.465
		50	1000	-0.646	-1.310	15.995
		80	1000	-0.683	-1.366	42.874
	50	30	1000	-0.448	-1.210	10.560
		50	1000	-0.451	-1.288	10.893
		80	1000	-0.524	-1.294	13.389
Complete Case[b]	20	30	1000	-0.461	-1.309	48.672

		50	1000	-0.397	-1.296	23.776
		80	1000	0.023	-1.362	99.621
	50	30	1000	-0.291	-1.196	13.004
		50	1000	-0.131	-1.288	16.084
		80	1000	0.323	-1.260	31.545

[a]Full, uncensored sample

[b]After censoring

[c]Bias was estimated only for samples where lognormal was the selected distribution because the R *EnvStats* package does not currently support gamma distributions for this approach

Table S15. Coverage of 95% approximate confidence intervals for the mean[a] from the simulation study for the moderately skewed lognormal distribution

Approach	n	Censoring Level (%)	Number Samples	95% CI Coverage
Full Sample[b]	20	30	1000	93.6
		50	1000	95.0
		80	1000	93.6
	50	30	1000	95.0
		50	1000	94.2
		80	1000	92.5
Full Sample Maximum Likelihood Estimation[b]	20	30	1000	93.4
		50	1000	94.9
		80	1000	93.5
	50	30	1000	96.0

		50	1000	94.9
		80	1000	94.4
Substitute CensoringLevel/2[c]	20	30	1000	90.3
		50	1000	87.5
		80	1000	75.7
	50	30	1000	90.8
		50	1000	76.8
		80	1000	63.8
Maximum Likelihood Estimation[c]	20	30	1000	92.0
		50	1000	92.8
		80	1000	77.2
	50	30	1000	94.4
		50	1000	91.1
		80	1000	75.4
Robust Regression on Order Statistics[c,d]	20	30	583	93.1
		50	458	93.0
		80	441	64.9
	50	30	560	96.1
		50	511	92.4
		80	424	67.7
Kaplan-Meier[c]	20	30	1000	89.1
		50	1000	67.1
		80	1000	3.8

	50	30	1000	86.4
		50	1000	42.3
		80	1000	0.2%
Complete Case[c]	20	30	1000	56.7
		50	1000	12.7
		80	1000	0.0
	50	30	1000	18.1
		50	1000	0.9
		80	1000	0.0

[a]Approximate confidence intervals of the form (sample mean) ± t_{n-1}(standard deviation)/√n

[b]Full uncensored sample

[c]After censoring

[d]Estimation only for samples where lognormal was the selected distribution because the R *EnvStats* package does not currently support gamma distributions for this approach

Table S16. Coverage of 95% approximate confidence intervals for the mean[a] from the simulation study for the highly skewed lognormal distribution

Approach	n	Censoring Level (%)	Number Samples	95% CI Coverage
Full Sample[b]	20	30	1000	76.0
		50	1000	74.3
		80	1000	74.2
	50	30	1000	79.8
		50	1000	80.0

		80	1000	79.1
Full Sample Maximum Likelihood Estimation[b]	20	30	1000	89.5
		50	1000	90.0
		80	1000	89.5
	50	30	1000	94.2
		50	1000	92.8
		80	1000	94.5
Substitute CensoringLevel/2[c]	20	30	1000	75.8
		50	1000	75.1
		80	1000	73.2
	50	30	1000	79.8
		50	1000	81.2
		80	1000	83.8
Maximum Likelihood Estimation[c]	20	30	1000	80.4
		50	1000	77.5
		80	1000	74.4
	50	30	1000	86.6
		50	1000	83.0
		80	1000	79.7
Robust Regression on Order Statistics[c,d]	20	30	583	85.1
		50	458	81.2
		80	441	80.3

	50	30	560	84.8
		50	511	86.4
		80	424	84.8
Kaplan-Meier[c]	20	30	1000	75.9
		50	1000	76.7
		80	1000	48.3
	50	30	1000	80.9
		50	1000	83.3
		80	1000	44.8
Complete Case[c]	20	30	1000	85.9
		50	1000	70.4
		80	1000	9.3
	50	30	1000	87.7
		50	1000	49.8
		80	1000	0.3

[a]Approximate confidence intervals of the form (sample mean) ± t_{n-1}(standard deviation)/√n

[b]Full uncensored sample

[c]After censoring

[d]Estimation only for samples where lognormal was the selected distribution because the R *EnvStats* package does not currently support gamma distributions for this approach

Table S17. Akaike Information Criterion (AIC) statistics for the group comparison

Data	n not censored	Normal	Lognormal	Gamma
Full sample[a]	85	-112.1	-205.2	-201.3
Censored at CCLV[b]	32	119.9	32.5	35.8

[a]AIC calculated using fitdist() function in R *fitdistrplus* package

[b]AIC calculated using fitdistcens() function in R *fitdistrplus* package

Table S18. Tests of group differences for resampled case study data for censoring at the MDL

Approach	Distribution	Stove Type 1		Stove Type 2		*p*-value
		n	Mean ± SD (ng/µL)	n	Mean ± SD (ng/µL)	
Uncensored						
Full sample	Normal	45	0.139 ± 0.156	40	0.079 ± 0.056	0.023[a]
Maximum Likelihood Estimation	Lognormal	45	0.145 ± 0.225	40	0.089 ± 0.115	0.103[b]
After censoring 16 observations[c]						
Substitute MDL/2	Normal	45	0.139 ± 0.156	40	0.078 ± 0.056	0.022[a]
Maximum Likelihood Estimation	Lognormal	45	0.142 ± 0.203	40	0.089 ± 0.109	0.083[b]

[a]Test for difference of means (i.e., that difference of means is zero)

[b]Test for the difference between the two groups expressed as the ratio of their geometric means

[c]Stove 1: 5 observations were censored; Stove 2: 11 observations were censored

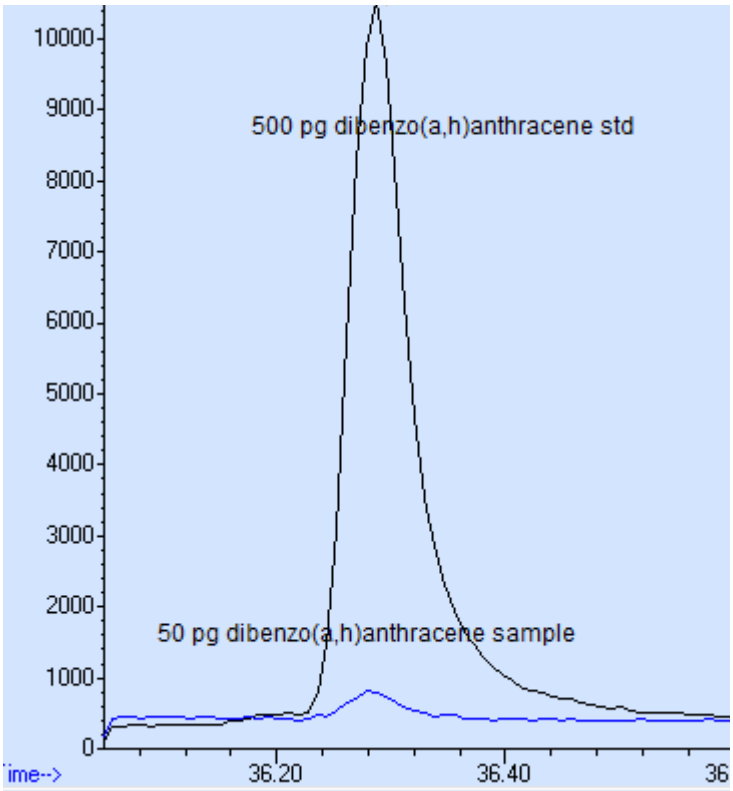

Figure S1. Example chromatogram of dibenzo(a,h)anthracene (DBA) standard (std; black line) and sample (blue line) with chromatographic retention time on the x-axis and response intensity on the y-axis.

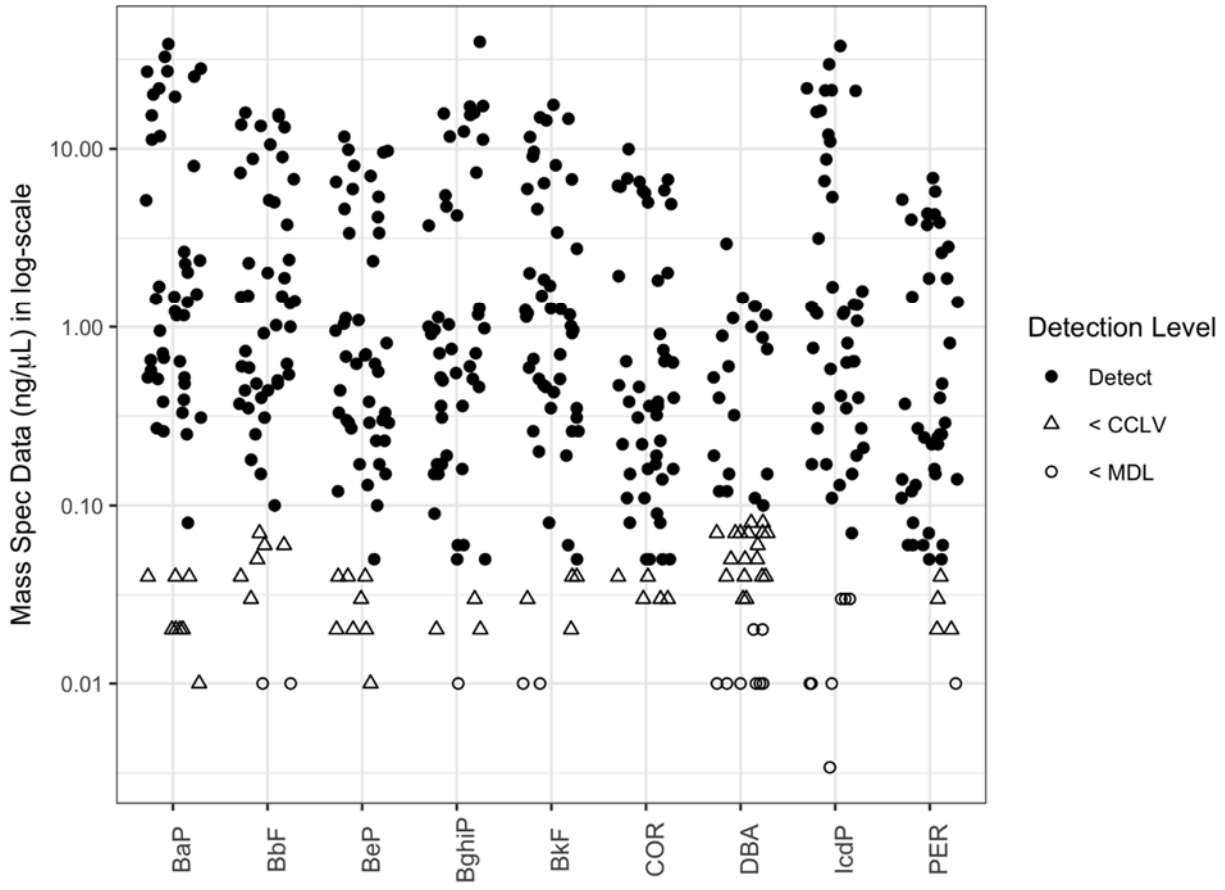

Figure S2. Concentration measurements below the method detection limits (MDL) are represented by open circles, and values between the MDL and the calibration curve lowest value (CCLV) are represented by open triangles. Concentration measurements above the calibration curve lowest value are labeled "Detect" and displayed as closed circles. Analytes are BaP = Benzo(a)pyrene, BbF = Benzo(b)fluoranthene, BeP = Benzo(e)pyrene, BghiP = Benzo(g,h,i)pyrene, BkF = Benzo(k)fluoranthene, COR = Coronene, DBA = Dibenzo[a,h]anthracene, IcdP = Indeno[1,2,3-c,d]pyrene, PER = Perylene. MDL and CCLV values and detection level category counts are given in Table S1 for each analyte.

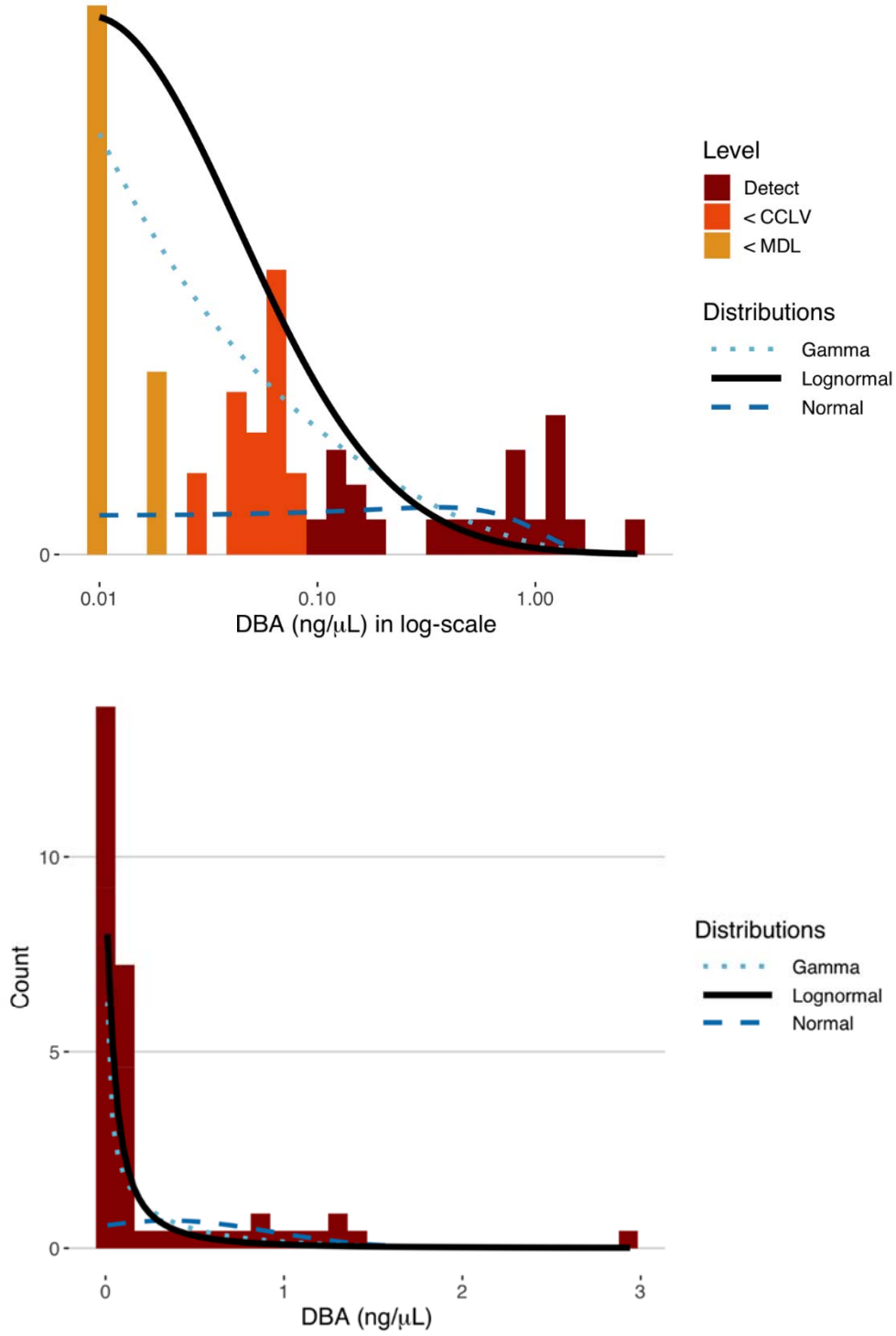

Figure S3. Histograms of the dibenzo[a,h]anthracene (DBA) samples (n=47) with log-scale x-axis (top panel) and original scale x-axis (bottom panel). Method Detection Limit (MDL) is 0.023 and Calibration Curve Lowest Value (CCLV) is 0.10.

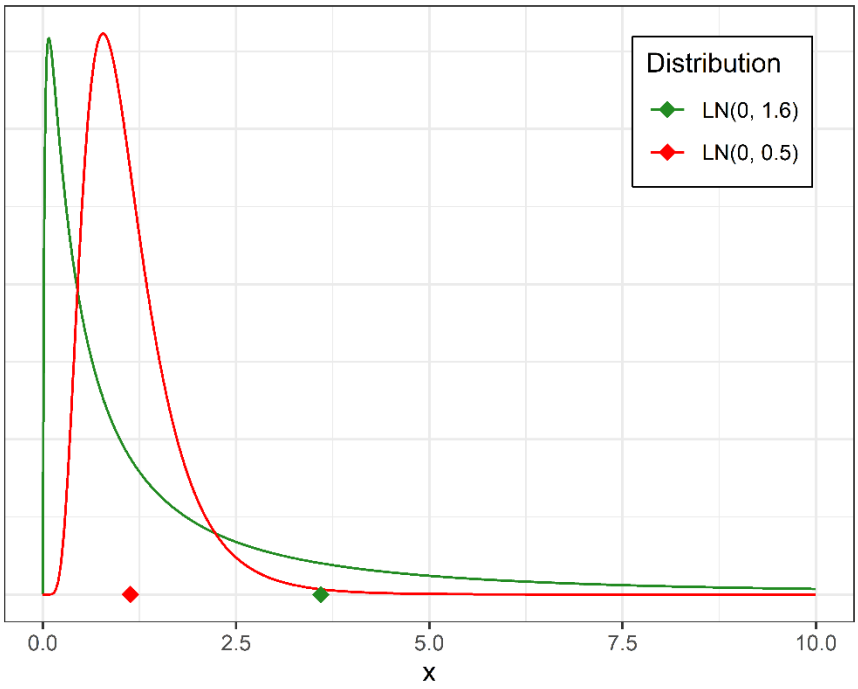

Figure S4. Lognormal distributions from the simulation study adjusted to have μlog=0, and therefore median of e^0=1, to contrast their means (indicated by diamonds) and variance characteristics. The distribution with σlog=1.6 has a larger mean than the distribution with σlog =0.5.

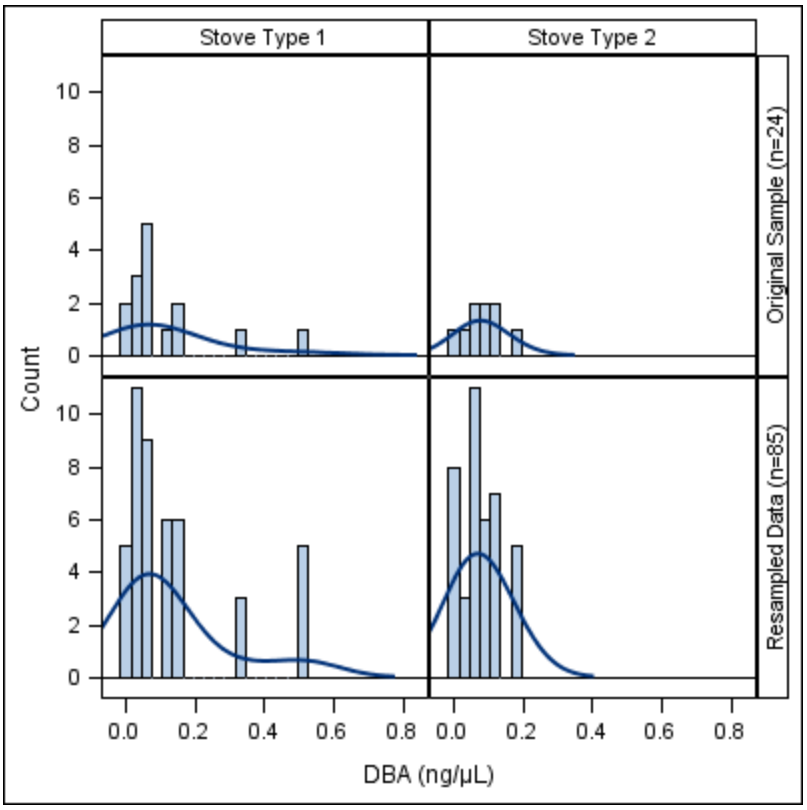

Figure S5. Histograms of original and resampled stove-specific data from the case study with kernel densities.

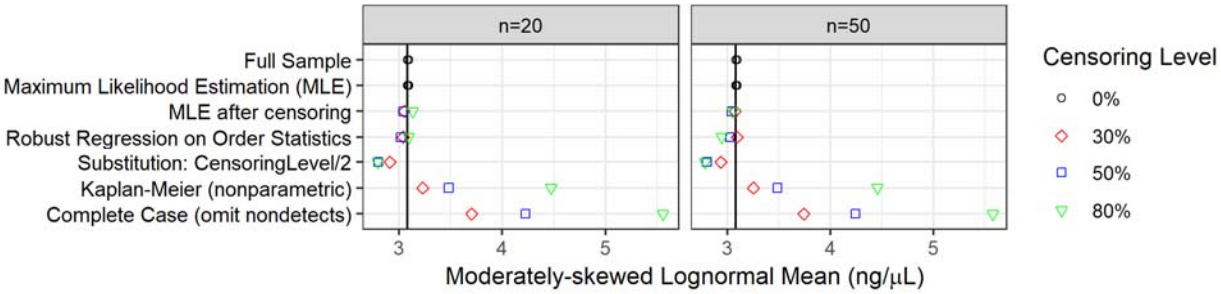

Figure S6. Bias in estimates of the mean from the simulation study for the moderately skewed distribution is the difference of the estimated means shown here and the true mean, which is indicated by the reference line. The true mean is approximately 3.08.

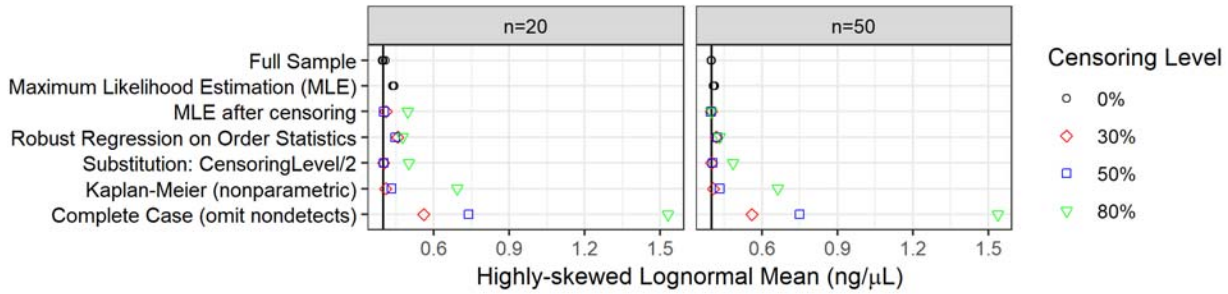

Figure S7. Bias in estimates of the mean from the simulation study for the highly skewed distribution is the difference of the estimated means shown here and the true mean, which is indicated by the reference line. The true mean is approximately 0.40.

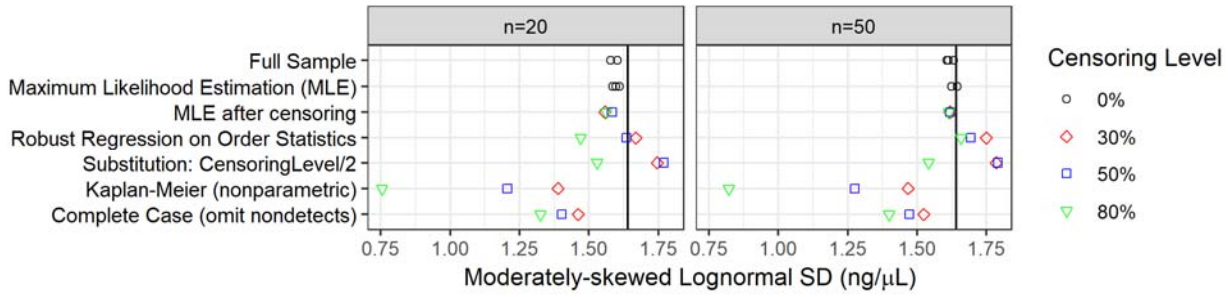

Figure S8. Bias in estimates of the standard deviation from the simulation study for the moderately skewed distribution is the difference of the estimated standard deviations shown here and the true standard deviation, which is indicated by the reference line. The true standard deviation is approximately 1.64.

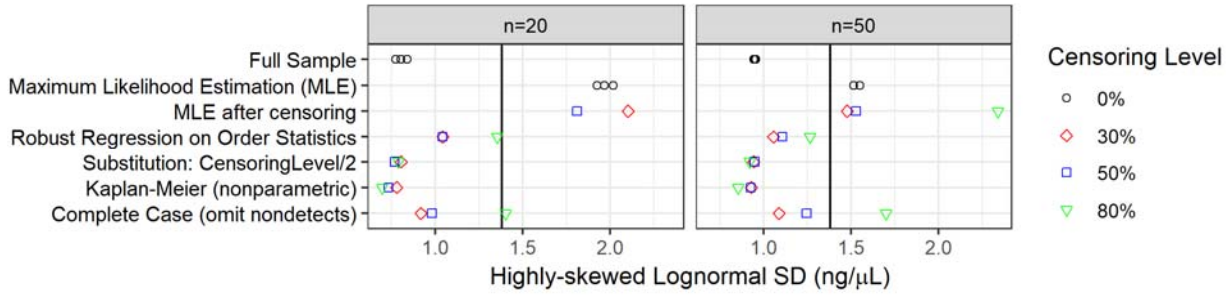

Figure S9. Bias in estimates of the standard deviation from the simulation study for the highly skewed distribution is the difference of the estimated standard deviations shown here and the true standard deviation, which is indicated by the reference line. The true standard deviation is approximately 1.38. Note: MLE after censoring estimate for n=20 and censoring level 80% is 269.4 and is omitted.

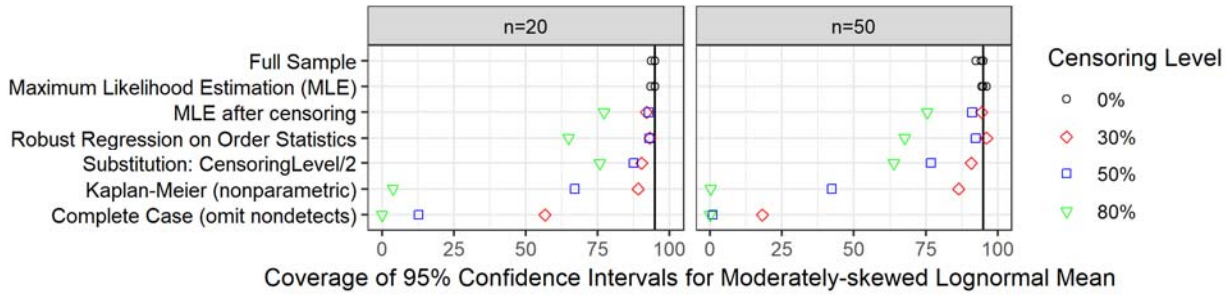

Figure S10. Coverage of 95% approximate confidence intervals of the form (sample mean) ± t_{n-1}(standard deviation)/√n for the mean from the simulation study for the moderately skewed distribution. Reference line indicates 95% coverage.

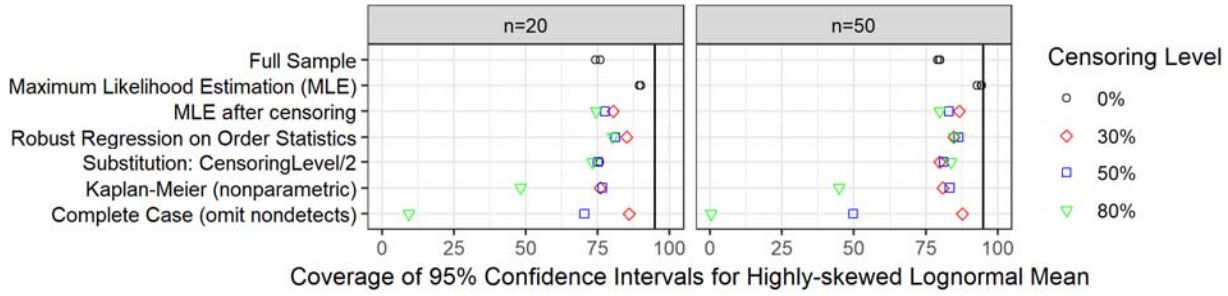

Figure S11. Coverage of 95% approximate confidence intervals of the form (sample mean) ± t_{n-1}(standard deviation)/√n for the mean from the simulation study for the highly skewed distribution. Reference line indicates 95% coverage.

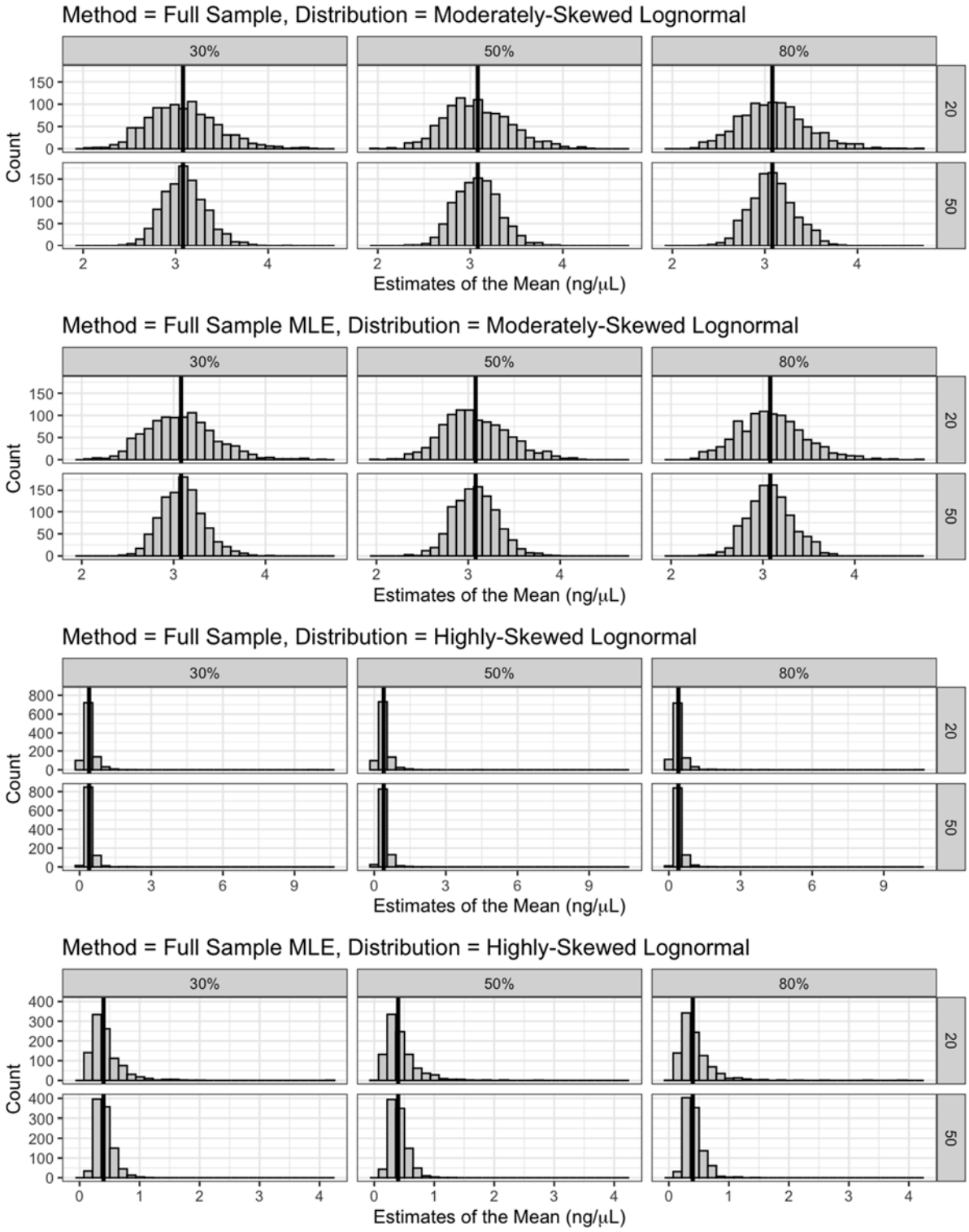

Figure S12. Histograms of simulation study results for full-sample and full sample maximum likelihood means for the moderately and highly skewed lognormal distributions, for censoring levels of 30%, 50%, and 80%, and for sample sizes 20 and 50.

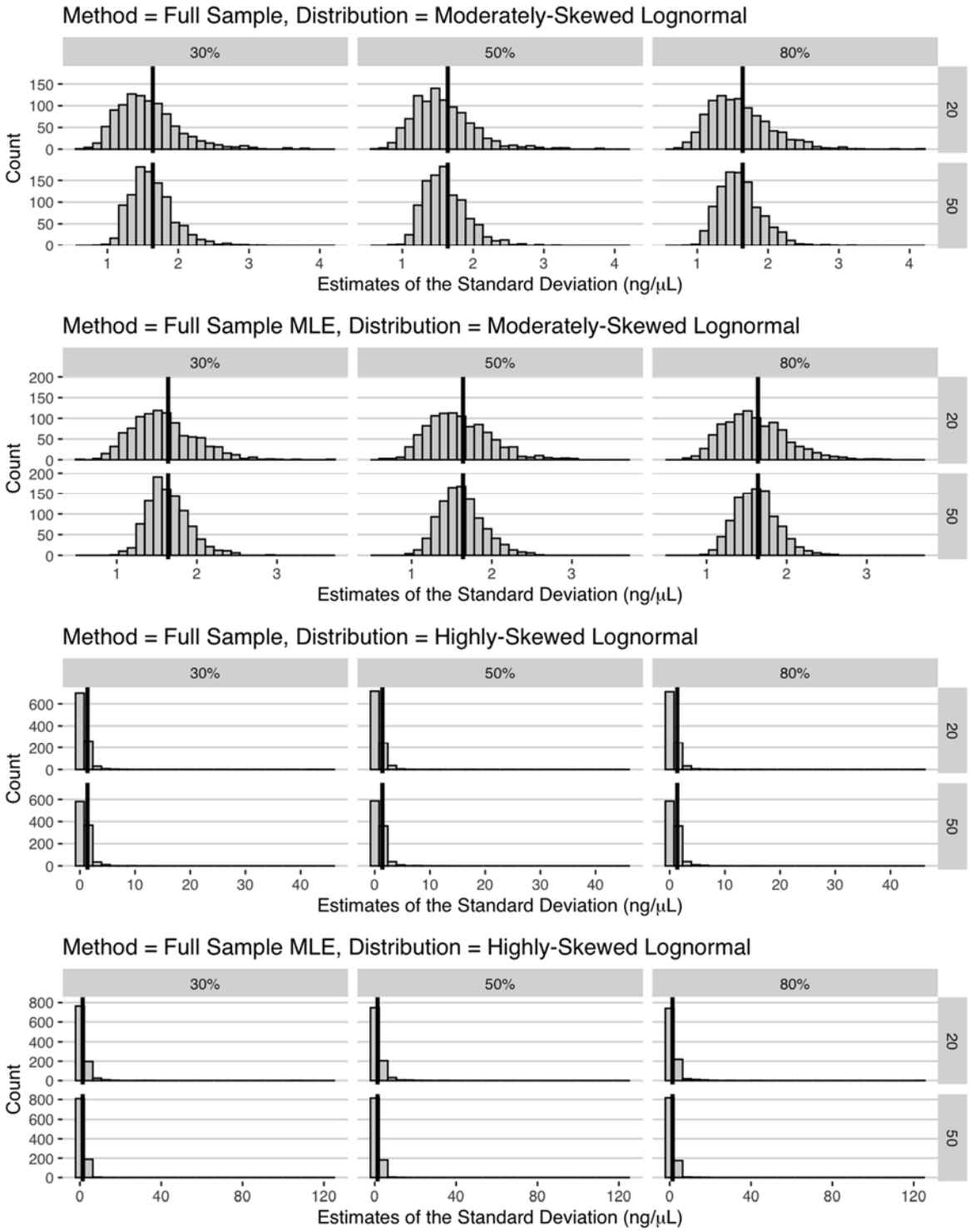

Figure S13. Histograms of simulation study results for full-sample and full sample maximum likelihood standard deviations for the moderately and highly skewed lognormal distributions, for censoring levels of 30%, 50%, and 80%, and for sample sizes 20 and 50.

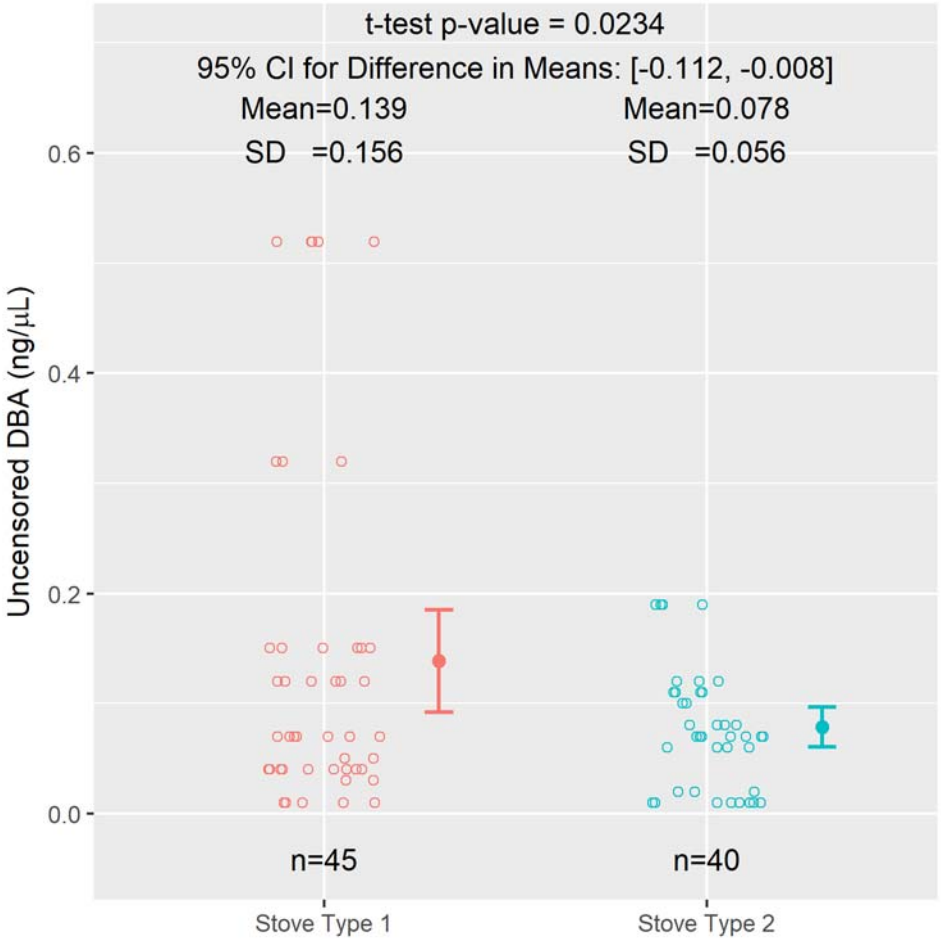

Figure S14. Scatter plots of uncensored resampled dibenzo[a,h]anthracene (DBA) values (ng/µL); also shown are the mean, 95% confidence interval for the mean, and standard deviation (SD) for each stove type and the *p*-value for difference of means t-test.

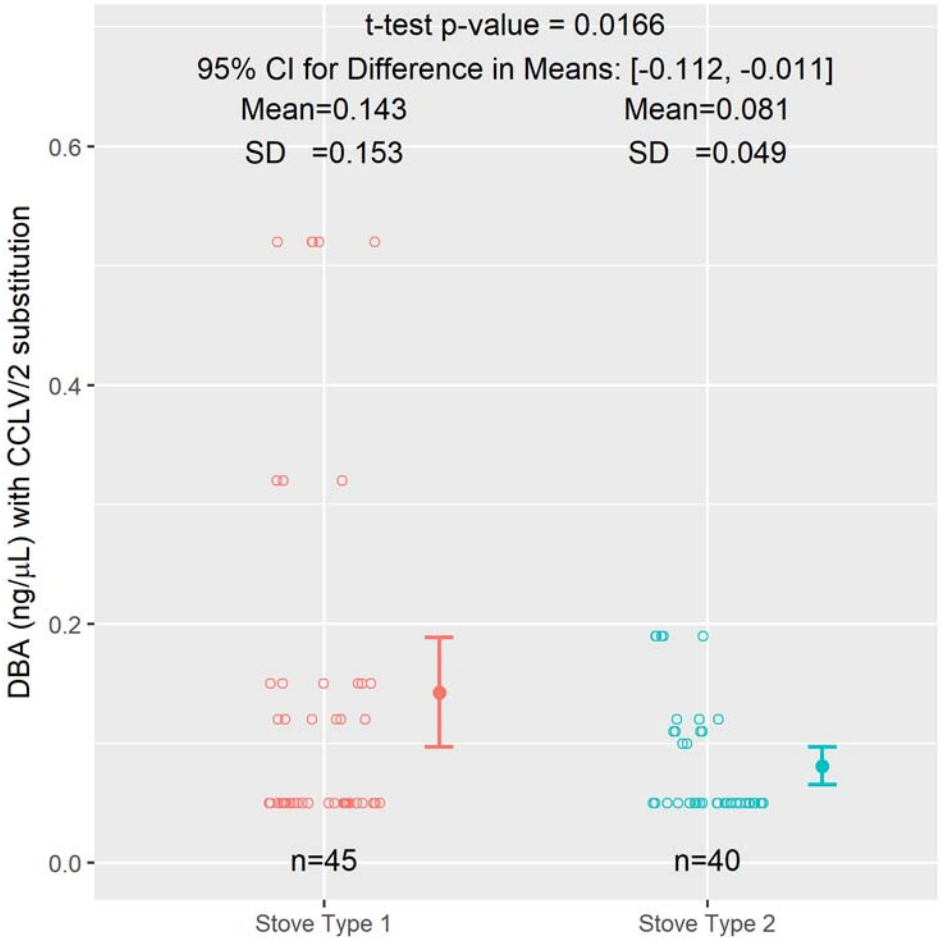

Figure S15. Scatter plots of resampled dibenzo[a,h]anthracene values (ng/µL) after censoring at the CCLV and substituting CCLV/2; also shown are the mean, 95% confidence interval for the mean, and standard deviation (SD) for each stove type and the *p*-value for difference of means t-test.