How to Handle Censored Industrial Hygiene Data

Technical Information Paper No. 55-039-0615

PURPOSE. To discuss the proper ways to handle censored industrial hygiene (IH) monitoring data when performing industrial hygiene exposure calculations.

REFERENCES.

a. DOD Industrial Hygiene Exposure Assessment Model, Report 2000-1, DOD Industrial Hygiene Working Group (DODIHW), January 2000.

b. Technical Manual NEHC-TM6290.91-2 Rev. B, Industrial Hygiene Field Operation Manual (IHFOM), Navy Environmental Health Center, Norfolk, Virginia, 23513-2617, March 1999 (see http://www.med.navy.mil/sites/nmcphc/industrial-hygiene/industrial-hygiene-field-operations-manual/Pages/default.aspx). [Note chapters have been updated continually since March 1999 revision.]

c. American Industrial Hygiene Association. 2006. *A Strategy for Assessing and Monitoring Occupational Exposures*, Third Edition. AIHA®, Falls Church, Virginia 22042. (AIHA® is a registered trademark of the American Industrial Hygiene Association.)

d. Paul Hewett and Gary H. Ganser. 2007. A Comparison of Several Methods for Analyzing Censored Data. *Ann Occup Hyg*; 51:611-632.

POINTS OF MAJOR INTEREST AND FACTS

Background

a. Industrial hygienists (IHs) conduct exposure sampling or exposure monitoring to determine worker exposures during the performance of tasks, processes, or operations. The sampling data is used to develop a probability density function (PDF) to describe the exposure profile for the population of workers or groups of workers with similar exposures (called a similar exposure group [SEG]). The PDF forms a curve with the exposure values on the x-axis and the probabilities of their occurrence on the y-axis. The exposures are usually reported as time weighted averages (TWA), but they can also be short term exposure limit (STEL) or ceiling/peak data. The PDFs come in two basic distribution curves: normal or lognormal. Inferential statistics are calculated for each of the PDFs to characterize the distribution and make judgments on whether the worker exposures in a SEG are acceptable or unacceptable.

b. In general, if the occupational exposure limit (OEL) is greater than 95 percent of the estimated population exposures ($X_{95\%}$) with some confidence factor, usually 95% (called the upper tolerance limit $(UTL)_{95\%,95\%}$), we say the exposures are acceptable; conversely, if the OEL is less than the $UTL_{95\%,95\%}$, we say the exposures are unacceptable. Similarly, if more than 5% of the exposures are greater than the OEL (exceedance fraction greater than 5%), we say the exposures are unacceptable (references a and c).

c. Data is considered censored when a sampling or exposure monitoring result is reported to be below the limit of detection (LOD), below the limit of quantification (LOQ), labeled as non-detected (ND), or above the maximum measurable concentration (> MMC). Censored data is a common outcome in IH exposure sampling or monitoring. The LOD and LOQ are not the same, but their names have been used interchangeably. Non-detected is another way a laboratory and/or IHs will report data that is below LOD and LOQ values. Data that is reported below a LOD/LOQ is considered left-censored. It is called left-censored because the values are plotted on the left hand side of the PDF curve. For left-censored data, the true value of the censored data lies between 0 and LOD/LOQ value. Censored data that is greater than (>) the MMC is considered right-censored data because values would be plotted on the right side of the PDF curve. The true value of right-censored data is unknown and difficult to estimate, so it is not discussed in this paper. Note if your data set contains both LOD and LOQ values, you should ask your laboratory to convert them all to LODs; this may remove some of the censored data from the data set (reference c).

d. Censored data sets can be divided into two categories: simple and complex. Simple-censored data sets have all the censored data at the low end of the data set, or PDF, with all the actual results located to the right of the censored data on the PDF curve. Complex-censored data sets have censored data spread throughout the data set, intermixed with the actual sample results. Complex-censored data sets occur when different sampling methods are used or different sample times and volumes are used with the same sampling methods to measure exposures in a SEG. This can result in multiple LOD/LOQ values, and in a complex censored data set you can have censored data with values higher than some of the actual data results (reference d).

Table 1. Simple and Complex Data Set Examples

Simple Censored	ND, ND, ND, 2.5, 2.8, 3.0, 3.6, 4.5, 5.2, 9.4, 15.8
Complex Censored	ND, ND 1.3, ND 2.7, 2.9, ND, 4.2, 5.9, 11.7, 14.9

e. Censored data sets are also generally rated based on the degree that the data set is censored.

(1) As a rule of thumb, data sets are divided into one of four categories outlined in Table 2 below (reference c). The degree of censoring for a data set is based on the percentage of data within the data set that is censored.

Table 2. Degrees of Censoring within a Data Set

Degree of Censoring	Percent (%) of Censoring Data Set
Low	Less than (<) 20 % of the data is censored
Medium	20% to 50% of the data in the data set censored
High	50% to 80% of the data in the data set censored
Severe	80% to 100% of the data in the data set censored.

(2) The degree of which a data set is censored will affect the accuracy of the PDF and its descriptive and inferential statistics. The more censored data in the data set, the more biased the statistical results. If the data set has only a low to medium degree of censoring, the PDF and its statistical values are useful; however, once a data set gets above 50% censored data, the resulting values become problematic. As a rule, if the data set is highly censored, the PDF and statistical values are still considered valid but should be reviewed closely when using them for making judgments. When a data set is severely censored, non-parametric methods must be used for the results to be valid. Non-parametric methods are beyond the scope of this TIP and therefore not discussed here.

f. Exposure assessment sampling data is used to calculate many IH statistics such as the $UTL_{95\%,95\%}$ and the exceedance fraction. These values are used to determine if exposures are acceptable or unacceptable. The question addressed in this paper is: how should censored data be used when making these IH-related calculations?

How Does Department of Defense Direct Personnel to Handle Censored IH Data?

a. The short answer is that the Department of Defense has chosen to substitute a value for censored data by dividing the LOD/LOQ value (<X) by the square root of 2 ($\sqrt{2}$) and substituting the result (X_c) for the censored data in the calculation. During the development of the Defense Occupational and Environmental Health Readiness System–Industrial Hygiene (DOEHRS-IH) database software, it was decided to use the Navy's Industrial Hygiene Field Operation Manual (IHFOM) requirement for handling censored data. The Navy requirement, outlined in chapter 4 of their IHFOM (reference b), requires the ($\sqrt{2}$) method. The DOEHRS-IH was programmed to convert any censored data (< X) to a substituted value by dividing the LOD/LOQ result by the $\sqrt{2}$ [(X_c=<X/$\sqrt{2}$)]. Therefore, when data is entered into DOEHRS–IH with a less than sign (<), the data is automatically converted using the $\sqrt{2}$ method described above. The result (X_c) is placed into the data table with a red asterisk. The converted data (X_c) is then used in all further calculations.

b. The approach chosen by the Department of Defense is not the only option for handling censored data. However, because the DOEHRS-IH method for handling censored data is automatic and written into the code, if you want to use other options for

handling censored data, you must do the calculations and subsequent judgment regarding the population's exposure outside of DOEHRS-IH.

c. Ideally, the method for handling censored data should be based on the individual situation. The initial method used depends on what one intends to calculate, how much accuracy and precision is required, and how much of the data is censored. The advantages and disadvantages of various methods are discussed below.

Acceptable Methods.

a. Substitution with Midpoints. Substitution with midpoints is where the censored data is replaced in the calculation by one of the following values: the LOD/LOQ value divided by 2, or the LOD/LOQ value divided by $\sqrt{2}$. The idea behind both of these substitution methods is that the substituted value is centrally located between 0 and the LOD/LOQ. The preference of division by $\sqrt{2}$ rather than 2 is that the resulting value is likely more depictive of the true PDF curve. The general recommendation for deciding which of these values you should use is based on the geometric standard deviation (GSD) for the data set. If the sample set has GSD that is less than 3, then divide the LOD/LOQ values by $\sqrt{2}$. If the GSD is greater than 3, the SEG should be reevaluated to ensure that all the personnel and operation/process/task truly belong in the SEG. If so, and the GSD is still greater than 3, then divide the LOD/LOQ values by 2. When the population of censored values in the sample (data) is below 50% (low to medium degrees of censoring), the accuracy of these methods is acceptable.

b. Log-Probit Regression. The Log-Probit Regression method uses log probability-units (probits) to generate estimates of the geometric means (GM) and GSD. The Log-Probit Regression method can be used for virtually any size data set. Mathematically, the method requires a minimum of at least three data points, and two of these data points must <u>not</u> be censored; however, this method is not recommended for small data sets (reference d). Basically, the method generates the best log-probit plot line that runs through the censored and non-censored data points and calculates regression coefficients which are then used to estimate the GM and GSD. These estimated GMs and GSDs are then used to generate the inference statistics such as the 95^{th} percentile point estimate, exceedance fraction, and $UTL_{95\%,95\%}$. This method can work with data sets having moderate degrees of censoring. This method is considered more accurate than simple substitution.

c. Maximum Likelihood Estimate. The third method is the Maximum Likelihood Estimate (MLE) method which is similar to the Log-Probit Regression. Basically, the method looks for the most likely fit for the log-probit lines that can be generated through all the data point, and the method selects the plot that provides the best estimate for the GM and GSD. The MLE method is beneficial because it can be used for data sets with a large number of censored data (up to 80% censored data sets). The MLE method

can be used for any size data sets, but some of the data must not be censored. Mathematically, it requires at least three data points, and two of these data points must not be censored (reference d). The method is considered the gold standard for data that is log-normally distributed. The MLE method is also considered more accurate (especially for lognormal data) than simple substitution.

d. Non-parametric Methods. Most of the methods discussed in this paper are useful for data sets that are up to 80% censored. To perform statistics with data sets that are severely censored (where > 80% censored), there are nonparametric methods available; however these methods are beyond the scope of this paper.

Methods to be Used Only after Careful Consideration.

a. Substituting Zero for the Censored Data. Because substituting the censored value with 0 has the general effect of driving the mean down and the standard deviation up, this method is not recommended. These changes in the mean and standard deviation will change the other statistical values that are used for making exposure judgments.

b. Removing the Censored Data. Remember that the goal of any exposure monitoring program is to determine the best representation of all the possible exposures that members of the SEG can experience. As discussed in the background section of this paper, we use the sampling data to develop the best exposure PDF that represents all the possible exposures that can occur in the SEG. The PDFs are defined by the statistical values used to construct them (such as GM and GSD). The removal of data from the data set will result in changes to the GM, GSD and other statistical values. Exposure data (results) should not be removed unless there is clear and documented reason as to why the data does not represent a valid exposure within the SEG or if the SEG was resampled. When the sampling method used had a high LOD/LOQ and resampling was performed (where the sampling method is changed or adjusted to have a lower LOD/LOQ that would ensure that the results would be above the LOD/LOQ), it is appropriate to remove the censored data and replace it with this new sampling data. Chapter 4 of the Navy IHFOM (reference b) discusses the resampling option. It recommends resampling the exposure where more than 50% of the values in a sampling run are censored by adjusting the sampling method and procedure to ensure that the results do not drop below the LOD/LOQ.

c. Substituting the LOD/LOQ Value. In this method, the LOD/LOQ value is substituted in place of the censored data. When the user calculates the normal distribution and log-normal distribution by substituting LOD/LOQ value for the censored data, the general effect would be to drive the mean up and the standard deviation down. These changes in the mean and standard deviation will change the other statistical values that are used for making exposure judgments.

Example.

a. Comparison of Censored Data Methods for Exposure Judgments. The following is a real example of the sample results collected to determine worker exposures to a metal. The sampling was performed on a welding operation over the course of several days. The data are listed in Table 3 below. There are 12 sample results, 4 of these results are censored. Thirty-three percent (33%) of the data is censored, and three (3) of the censored data results are at the lower left side of the distribution. One censored result is located near the middle of the data set; therefore, this data set is considered a moderately censored complex data set. Table 4 shows that each method leads to a different inferential statistics results. Note the $UTL_{95\%,95\%}$ is calculated using n'=8 and not n=12. (Note that n' is the total number of samples collected minus the number of samples collected that were censored. See paragraph b, Recommendation on Calculating Confident Limits Using Censored Data, below, for discussion.)

Table 3. Example Sample Results (Data), Sample Results for a Metal in mg/m^3 (OEL = 0.01 mg/m^3)

Data	<0.0011	0.0023	0.0013	<0.0011	0.0016	0.012
Censored	y	n	n	y	n	n
Data	0.0012	<0.0011	0.0045	<0.003	0.005	0.007
Censored	n	y	n	y	n	n

Note:
mg/m^3 = milligrams per cubic meter

Table 4. Comparison of Different Methods of Handling Censored Data (OEL= 0.01 mg/m^3)

Censored Data Method	GM	GSD	$X_{95\%}$	$UTL_{95\%,95\%}$ (n' = 8)
Removal	0.0032	2.34	0.0129	0.0478
LOD	0.0024	2.29	0.0095	0.0342
LOD/2	0.0019	2.82	0.0107	0.0529
LOD/√2	0.0022	2.52	0.0099	0.0412
Log-Probit	0.002	3.08	0.0129	0.073
MLH	0.0019	2.88	0.108	0.055

Note:
MLH – maximum likelihood

b. Recommendation on Calculating Confidence Limits Using Censored Data. When you calculate confidence limits (CL), such as the 95% confident limit for the 95% percentile or $UTL_{95\%,95\%}$ with a data set that has censored data, the value of **n** or the

number of data points (or samples) used in the calculation should be modified by subtracting the number of censored data points or samples (called **m**) from n. The total data points minus the total number of censored data points is sometimes referred to as **n** prime or simply **n´**; therefore **n´ = n-m**. The **n´** is substituted for **n** in the calculation of the CL. The reason that it is recommended to use **n´** in calculating CLs is because the censored data and methods used to deal with censored data discussed in this paper do not represent the true values, and if you used an unmodified **n** the resulting CL would misrepresent the true CL. The difference between the calculated CLs using **n** and using **n´** can be significant (depending of sample size and number of censored data points). The CL calculated using **n** can be as much as half the value as using **n´**, which can introduce a significant bias in the decision or judgment on acceptable versus unacceptable exposure. Therefore, when using any statistical software or Excel® make sure that **n´** is used when calculating CLs (reference c). Table 5 shows the difference in the $UTL_{95\%,95\%}$ if n=12 was used (incorrect) rather than n'=8 (correct). Note that DOEHRS-IH and AIHA free statistical software does not support this process. In most cases when you are dealing with low- to medium-censored data set, this may not significantly affect your decision; however, if you are dealing with higher degrees of censored data sets, this procedure for calculating the CLs should be used. (Excel® is a registered trademark of Microsoft Corporation.)

Table 5. Comparison of UCL 95 Calculated of Using All the Results (n) and Only the Non-censored Results (n´)

Censored Data Method	Total Number of Sample = n	$UTL_{95\%,95\%}$		Number of Non-Censored Samples = n´	$UTL_{95\%,95\%}$
LOD	12	0.0235		8	0.0342
LOD/2	12	0.0331		8	0.0529
LOD/√2	12	0.0272		8	0.0412
Log-Probit	12	0.044		8	0.073
MLH	12	0.0341		8	0.055

Conclusions.

a. There are several methods for handling censored data. All the methods will affect the calculations of the statistical values which are used to characterize the SEG's exposure. Therefore, the resulting exposure judgments will be biased to some degree and could result in potentially different decisions. Note that when selecting a method for handling censored data, there is a tradeoff between simplicity of use versus accuracy and precision.

b. In most cases that occur in IH, the substitution for the censored data with the LOD/LOQ divided by √2 will provide adequate estimates for the statistical values used for making SEG exposure judgments. The user should be aware of the bias produced

by the method(s) they select and should consider using multiple methods when the exposure acceptability judgments are close to fully assessing workers exposures in a SEG. The use of multiple methods will allow the user to see all the potential assessment outcomes and determine the appropriate assessment that meets their goals.

Recommendations.

a. Do not remove censored data from an exposure assessment data set unless it can be determined and documented that the data is not representative of the exposure.

b. Continue to use the substitution method where a value is generated by dividing the LOD/LOQ value by the $\sqrt{2}$. If the data set had GSD greater than 3, consider dividing the LOD/LOQ value by 2, but first reevaluate the SEG to ensure all members truly belong in the SEG.

c. When the statistical values used to make a judgment are borderline between acceptable and unacceptable or when dealing with highly censored data sets (>50% to 80%), consider using the Log-Probit Regression or one of the MLE techniques to provide better estimates of the GM and GSD.

Point of Contact.

The USAPHC Army Institute of Public Health point of contact is the Industrial Hygiene Field Services Program, commercial: 410-436-3118, or DSN: 584-3118.

Prepared by: Industrial Hygiene Field Services Program
Dated: June 2015