Censoring Data - An Example

Should uncensored data be censored? A simple example shows that censoring data should, in general, not be used, or used only with extreme caution, and with full knowledge of its consequence.

Frequently, laboratories will report radionuclide data at levels below the minimum detectable concentration (MDC), and even occasionally data with negative values. How should this data be interpreted, and more importantly, how should it be used to calculate a variety of parametric or non-parametric statistics?

Although it is true that negative radionuclide concentrations cannot exist, a negative value reported by the laboratory does has value and meaning. This is because a laboratory does not directly measure concentrations. It measures the number of radioactive particles detected during a fixed count period from a sample, and that count will include the instrument background. This is the gross count. The instrument background count is then subtracted from the gross count to give the net sample count. The net count can be negative under certain conditions since background count at any one time is stochastic and therefore variable. See below. This net count rate is then used to calculate a concentration using sample mass or volume, count time, detection efficiencies, geometric factors, flow rates, unit conversions, etc.

All radionuclide analysis involves counting the number of radioactive decays (either gammas, alphas or betas) emitted by the sample per unit time within a low-background laboratory counter. Even though counters are shielded to minimize any extraneous radiation entering from the outside or from within the equipment itself, there will always be a low level of radioactive particles detected even with no sample present or a blank sample present. This is known as the instrument background. For example, let us say the instrument background is measured at 10 counts per minute (cpm). The MDC expressed in cpm will be $3 + (2 \times 1.645 \times (2 \times 10)^{1/2}) = 17.7$ cpm. Now let's count a sample that is <u>not contaminated</u> 10 separate times. We might imagine that with a non-contaminated sample we would get 10 cpm each time. However, since we are counting background <u>plus</u> the sample (gross count) and since background (and the sample) is variable, and will fluctuate statistically during each of the counting periods, we may get the following gross counts.

10, 11, 12, 9, 9, 10, 7, 13, 11, 8.

The average of these gross counts is 10. Subtracting the single background count of 10 cpm and ranking, we get the following net counts.

-3, -2, -1, -1, 0, 0, 1, 1, 2, 3

Note that some are negative net counts, and all are less than the MDC of 17.7 cpm, therefore all are considered non-censored non-detects. The simplest parametric statistic for this data set is the arithmetic mean which is calculated to be 0 cpm, which correctly confirms that the

sample is not contaminated. However, if we were to dismiss the negative net counts as meaningless, the mean of the left-censored data set …

0, 0, 0, 0, 0, 0, 0, 1, 1, 2, 3

… would be 0.7 cpm, which would incorrectly imply the sample slightly exceeded background. However, the error introduced is relatively minor and the average is considerably less than the MDC of 17.7 cpm and therefore the sample would be judged to be not contaminated.

If we were to left-censor the data set even more, dismiss all data less than the MDC, and substitute all data with the MDC value, the left-censored data set would become …

17.7, 17.7, 17.7, 17.7, 17.7, 17.7, 17.7, 17.7, 17.7, 17.7

… with an average of 17.7 cpm. The non-contaminated sample would therefore be declared contaminated at 17.7 cpm.

Likewise, left-censoring the entire data set and substituting the data set with MDC/2 values would show the sample to exceed background by 8.85 cpm (though still less than the MDC)[1].

Clearly valuable information would have been lost by censoring data and parametric or non-parametric statistics calculated based on this censored data set could give incorrect quantitative results.

This example illustrated the problem when counting one sample multiple times. A data set of multiple samples, counted once, with many <MDC values and some negative values, is obviously a somewhat different situation, however the philosophy remains the same. Censoring data should only be used with extreme caution with full knowledge of its consequences.

[1] Note that the practice of substituting <MDC data with MDC, MDC/2, MDC/$2^{0.5}$ or 0 is a quick and easy solution when laboratory data is actually reported as a qualitative <MDC, not when the laboratory reports quantitative values that happen to be <MDC.