Practical Stats LLC
Statistics, down to earth

# Practical Stats Webinar: NADA2

## Questions submitted, with answers

1. For AIC, you didn't substitute for censored data, but did you have to indicate in your raw data that those values were below detection, and give the detection limit?

Yes, that information is crucial. The indicator and the detection limit comprise the information that is available in nondetects. These are what are used to modify the percentiles of the detected values -- if 30% of the dataset are below the single detection limit of 1, then the lowest detect (say at 1.4) will be just above the 30th percentile. In our Nondetects And Data Analysis online course I have also explained how to go to 'plan B' when you don't have the detection limit value (hopefully this only occurs for old data) and nondetects are just listed as "ND" or "trace" or something similar. You can use the lowest detected value in that case, but it will be less informative and slightly biased upward than if you knew the actual detection limit value.

2. Will NADA2 be a standalone package, or an "add-on" for NADA (i.e. will it be necessary to load both)?

Neither. NADA will be an addon to NADA2. In R terminology, NADA as well as 14 other packages will be 'dependencies' to NADA2 -- NADA2 calls functions that are provided in one of those 15 packages. This is because NADA2 includes a wide range of methods, parametric and nonparametric and permutation and regression and multivariate procedures, etc. Other R packages call a similar number of dependencies. The best way to handle this is to write a script that loads those packages needed with one command. In the Nondetects And Data Analysis online course I've written an example script to load all the libraries needed with a single command. Only takes a second. People could take that script and add their own packages that they often use. Then start each R session by running that script and you're ready to go. There's no getting around that R is modular software. You only use what you need.

3. Will Quantile Regression be available in NADA2?

No, quantile regression is available in its own R package, quantreg. It is very useful for using regression to predict how quantiles (percentiles) of the data change as a function of the X variable(s), rather than predicting the mean as in ordinary regression.

4. Do you have examples for comparing censored data to standards?

Yes, I covered that in two of our newsletters. My recommendation is that you use permutation tests so that no distribution needs to be assumed. Censored data can easily be done in a permutation format and is done that way in the NADA2 package. See our

August and October 2018 newsletters in the News archive page at practicalstats.com for an example of doing the equivalent of a t-test comparing censored data to standards. The News Archive contains free newsletters that I've sent out since about 2006. They give short introductions to many topics in environmental statistics, the topics I've covered in the courses I've taught over the years. More detail is available in the two courses I still teach online, the NADA course and Applied Environmental Statistics.

**5. How do you adapt these left-censored data routines to interval censored data (nondetects and exceedances)? How would you draw a boxplot with multiple nondetect limits and exceedances?**

Second question first. Boxplots with multiple detection limits are drawn using the cboxplot function in NADA2. The standard way is to draw a horizontal line across the plot at the highest detection limit. Everything below that is grayed out, or dashed, or omitted, any of which communicates that what is below must be estimated (with a reasonable model). Percentiles for detected data below the highest DL must be estimated with a model of some sort that incorporates both detects and nondetects below the highest limit. I've relied upon the ROS (Regression on Order Statistics) method to do that because it doesn't depend strongly on the distribution chosen as the model. More detail on the procedure is found in my 2012 book, *Statistics for Censored Environmental Data Using Minitab and R*.

Several (but not all) of the routines in NADA2 will accept interval-censored data as well as the indicator format I used in the webinar. The first two columns of data input are the low and high ends of the interval within which the data lie, so (0,1) for a <1 and (1, 3) for data between the DL of 1 and the quantitation limit of 3, for example. The mathematics of interval censoring is just a combination of left-censored and right-censored equations so that the interval is modeled. Interval format input is allowed in an assortment of nonparametric methods as well as parametric. Limitations in the code for regression methods make things like computing residuals from a regression model problematic, so not everything in NADA2 can be done with either format. Adding more interval-censoring input to NADA2 routines is one of my primary goals to complete before NADA2 is released. A second goal is to finish writing the help files, which must be completed before an R package can be published.

**6. Thank you for your great webinars! When working with many thousands of measurements in machine learning models for regression, I have replaced censored values in the response variable with an estimated value based on maximum likelihood estimation. I have done this when the proportion of non-detects is relatively low, and tend to switch to a classification model if the proportion of non-detects in the response variable is large. As far as I know, there are no ML methods that can handle the non-detects directly as the methods in this webinar do. For regression ML methods, do you recommend a different treatment for non-detects besides MLE? Or are there ML regression methods which can take the non-detects into consideration?**

The regression method in NADA2's cencorreg function is maximum likelihood estimation. It estimates the slopes and intercept without estimating values, so if that is what you mean they do estimate them directly while taking nondetects into consideration.  I suggest that you look at the free webinar "The Mystery of Nondetects" on our Training Site, practicalstats.teachable.com.  It describes how MLE but also the other methods used operate.  If one is planning to estimate individual values, MLE is one of the best ways to do it.  However, the problem remains that if there were 10 (let's say) values at <1 there is no specific way to assign any given estimated value from MLE (which all differ) to any particular one of the 10 <1s.  So to see how this affects the results it is best to do this in a permutation approach, doing it many times with many different assignments to different <1 observations to see how variable the results are.  Another issue with ML in environmental studies is the people often assume normality, which allows a portion of the distribution (and estimated values) to fall below zero.  This is unrealistic and also produces estimates of the mean that are biased low.  That is discussed in more detail in our NADA online course.