Analyzing Values Below the Method Detection Limit Using R

Carolyn Huston
Simon Fraser University

Hosted by the Bulkley Valley Research Centre

February 2008

SFU

Schedule For The Day

8:30-9:00am	Introduction to Topics and Warmup
9:00-9:40am	Getting Started With R
9:40-10am	Computer Exercises
10:00-10:15am	Coffee Break
10:15-11:10am	More on R Basics
11:10-12pm	Computer Exercises
12:00-1:00pm	Lunch
1:00-1:15pm	Introduction to Non-Detects
1:15-2:00pm	Analyzing Data With Non-Detects 1: Graphing
2:00-2:45pm	Computer Exercises
2:45-3:00pm	Coffee Break
3:00-3:40pm	Analyzing Data With Non-Detects 2: Summary Stats
3:40-4:20pm	Computer Exercises
4:20-4:30pm	Wrap-up

Word Association

Let's play a word association game! I am going to show you some words/phrases below. As they are revealed, feel free to express some words or feelings that you associate with these concepts

- Statistics
- Learning new statistical software!

Learning About You

- How many of you are involved with scientific research in some way?
- What are some of the goals of good science/scientific method/your own research?

Learning About You

Hopefully we came up with some of the following points (and more)

- Replicability!
- Amenable for peer review
- Future research/extensions possible
- Other?

Why R?

How many of you have ever done the following?

Spent days/weeks/months/years collecting data, processing it, and finally entering it in the computer. After struggling with Excel/SPSS/etc. for awhile you finally figure out the 'correct' analysis, and get a p-value that you can report. A week/month/year after generating the p-value somebody asks a question about the results, so you go back to your data and think "huh, I wonder how I did that?"

Why R?

Issues like the above speak to a problem that occurs fairly frequently in scientific research with regards to replicability of numerical analysis. A similarly designed study might yield similar data and be replicable in terms of recreating similar data. It is equally important, though, that the statistical analysis itself can also be repeated and explained.

The R program for statistical computing helps to create easily replicable analyses once data has been collected

Why R?

In their product quality testing, many large corporations (eg. *SC Johnson*) require that the data analysis for product testing be instantly and completely replicable. They require both the data, and the script that was used to generate any statistical results to be submitted on record. This means that spreadsheet/point and click programs such as Excel/SPSS/JMP are not considered acceptable. This is because in the event of government scrutiny, or a lawsuit, their safety testing procedures must be completely transparent such that opposing experts can reproduce results, and search for errors in the methodology used to create them.

- Aside on *Oust*®

Why R?

The requirements discussed above essentially restrict statistical analyses to the use of *scripting* statistical software suites such as SAS, R, or S-Plus. Of these choices, I am advocating and teaching R software methods for use in water quality data analysis because it is free, and also because it has an add on package that is intended specifically for the analysis of data containing observations below the method detection limit (more on this later)

Why R?

Up to now we have basically discussed 2 of the 3 main reasons that I think working with R for data analysis is a step forward:

- It provides consistent answers that are easy to replicate, even years later
- It is also easy for work colleagues and peers to look at, replicate, and edit. Peer review and oversight is beneficial.

Why R?

This brings us to one more reason I think R is a useful program:

- It is easy to develop or extend a method or analysis in R

There were numerous extensions made to the base software in R during the writing of the guidance document, and we will play with some of these later in the day.

Now that I have you all motivated and excited.....

Let's get started learning R!!!!!!