The Nitty Gritty!!

Topics for Discussion

- Data Storage
- NADA and Censored Data
- Censored Boxplots
- Substitution

Data Storage

In a very sneaky way, we have already addressed some of the most important data storage issues! If you were/are able to create a data format using the `rep()` and `seq()` functions, then you are already well on your way! The type of data formats that you can create using these functions tend to be amenable to analysis in R (and most other statistical software)

How does the way you are currently storing data relate to how you would structure the same data sets in R?

Data Storage

Assuming that data you are using is in a format that is R friendly, there is one more really important aspect to data storage. How to store censor values is very important! While there are many ways of representing this type of data, only a few of them are commonly used in statistical software programs.

Data Storage

One of the most commonly seen ways to store data is seen below:

Table: Values of <0.001, <0.002 and 0.003 represented using a "negative" number.

-0.001
-0.002
0.003

This is not a very good method though... problems?

Data Storage

Another method

Table: Values of <0.001, <0.002 and 0.003 represented using less thans.

<0.001
<0.002
0.003

Also not a great method... why?

Data Storage

Another method... again

Table: Values of $<$0.001, $<$0.002 and 0.003 represented using TRUE and FALSE. A TRUE indicates a censored observation.

Value	Indicator
0.001	TRUE
0.002	TRUE
0.003	FALSE

Look familiar? .... It is almost like I have a plan!!

Data Storage

Last method...

Table: Values of $<$0.001, $<$0.002 and 0.003 represented using interval endpoints.

Start	End
0	0.001
0	0.002
0.003	0.003

Not as common in NADA, but used in the guidance document when conducting paired comparisons.

Censored Data

Before lunch, we talked about objects and how different functions can be used on different types of objects. We focused specifically on two different types of data objects: vectors, and data frames. When we load the NADA package in R, R allows us to create a *censored* data object. Such an object consists of data in the form of the table containing the TRUE/FALSE indicators.
Most of the functions from NADA will require you enter the data in this format.

Plotting

- One of the most effective ways to communicate numeric information is in the form of plots. I recommend plotting your data whenever possible!!
- Features of data can usually be observed more easily on a graph. Just think how much easier it is to look at plots and pictures than it is to look at a huge data table!!
- This is because the human eye has evolved to identify things visually - think about facial recognition.

Plotting

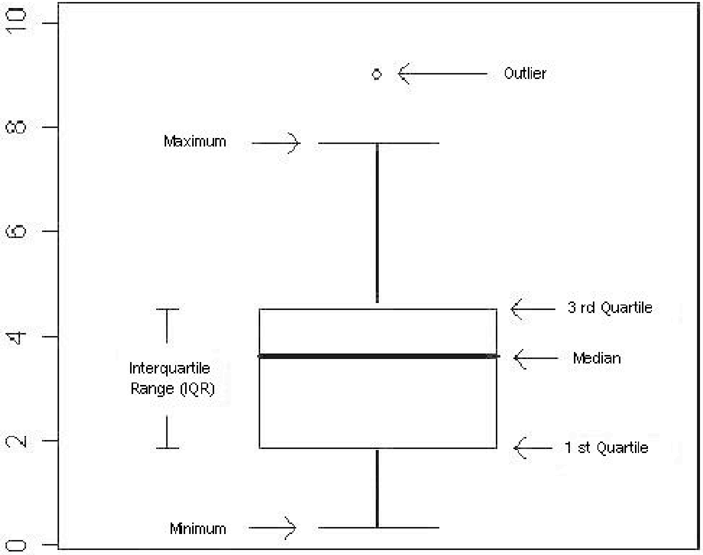
Granted, it can seem like science, and a lot of other research areas are pretty obsessed with creating and reporting as many p-values as possible, but before you discount graphs, think about the following:

How easy is it to remember people, compared with remembering people's license plate numbers? Which is a better identifier?

Plotting:Boxplots

One of my favourite kinds of plot to use when looking at data is a boxplot. There is a nice big section on boxplots in the guidance document. I will cover a few key points here.

Boxplots are drawn based on the quantiles of the data. This means we can see the minimum, 25^{th} percentile, median (middle), 75^{th} percentile, and maximum of the data all in one easy plot. Boxplots will also often identify unusual observations.


0
2
4
6
8
10
Outlier
Maximum
3 rd Quartile
Median
Interquartile
Range (IQR)
1 st Quartile
Minimum

Plotting:Boxplots

Things to comment on when talking about boxplots

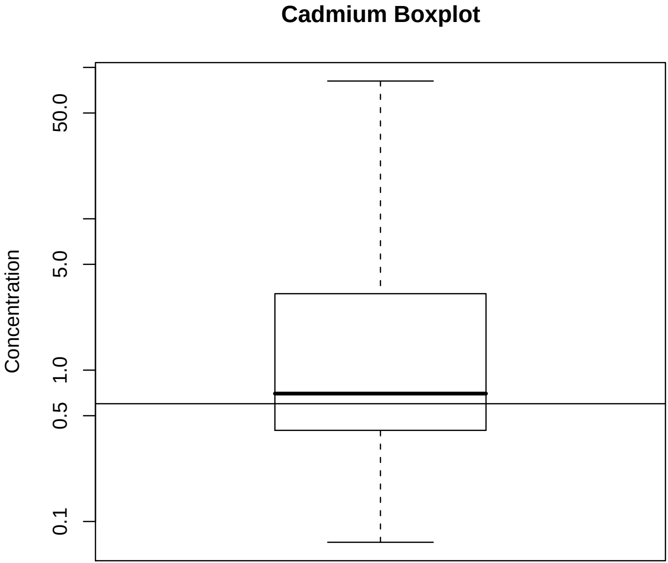
- Shape
- Typical Values
- Spread
- Outliers

Censored Boxplots

When we have a data set that has left censor values, we can draw a boxplot in the same way as I have shown you above. The big difference is that we can't really comment on the shape of the boxplot below the value of our highest censor limit.

The NADA package has a function called `cenboxplot()` that draws a *censored boxplot*. A censored boxplot is the same as a regular boxplot, except there is a horizontal line representing the highest censor limit. We can accurately comment on the attributes above this line, but not below.

Censored Boxplot



Drawing Your Own Censored Boxplot

To draw the censored boxplot shown on the previous slide, you can use the following commands in R: You don't have to do it now, it will be in the Exercises for this section.

1. Load the data into R using the following command (this data set is automatically loaded with NADA)

   ```
   >data(Cadmium)
   ```

2. Then use the `cenboxplot()` command in either of the following ways:

   ```
    >with(Cadmium,cenboxplot(obs=Cd,cen=CdCen,
    main="Cadmium Boxplot",ylab="Concentration",log=TRUE))
   ```

   or

   ```
   >cenboxplot(obs=Cadmium$Cd,cen=Cadmium$CdCen,
    main="Cadmium Boxplot",ylab="Concentration")
   ```

   Note the Argument log=TRUE

Why not substitution?

At this point, some of you who have worked with data containing left censoring might be saying to yourself ‘We could do all this fancy censoring stuff, OR we could just substitute the values below the detection limit with some sensible value!’

How many of you have ever substituted for values below the detection limit, or seen it done? How about deleting values below the detection limit?

How did you determine what a sensible value for substitution was?

Boxplots Based on Substituted Values

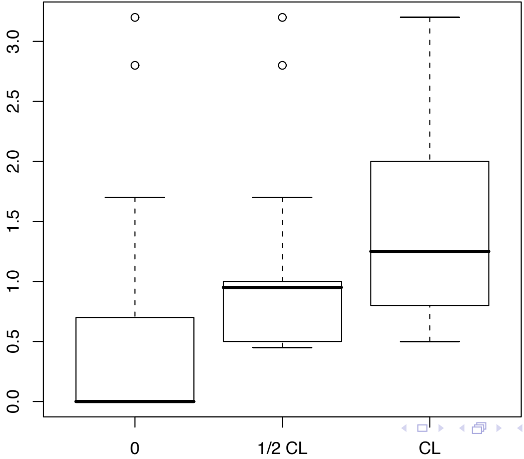
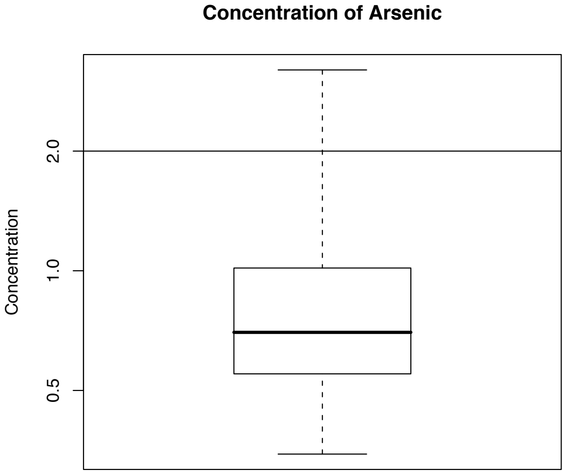


Figure: Boxplots after substituting the censoring limits for 0, 1/2 CL and CL.

Correct (Censored) Boxplot

Figure: Boxplot of arsenic concentration


Concentration of Arsenic
Concentration