LAST SECTION!!!

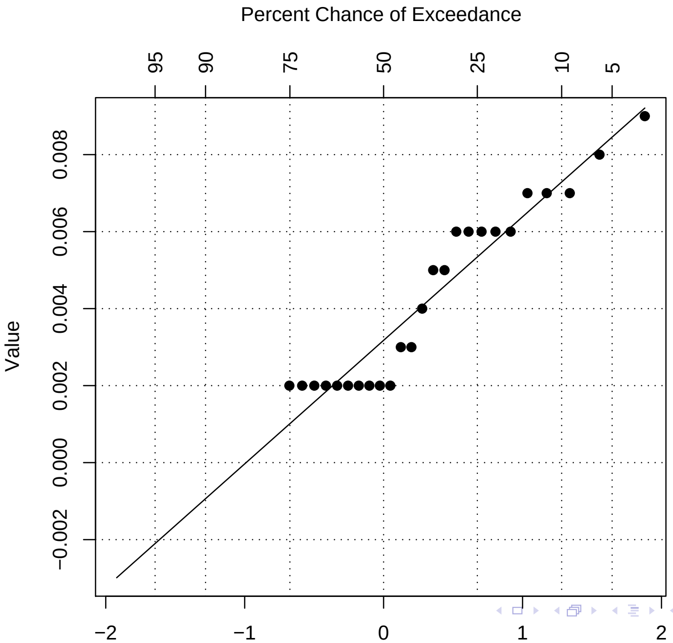
Some Topics

- **Probability Plotting**
  - Normal Distributions
  - Lognormal Distributions
- **Statistics and Parameters**
- **Approaches to Censor Data**
  - Deletion (BAD!)
  - Substitution (BAD!)
  - Parametric Methods
  - Non-Parametric Methods
  - Semi-Parametric Methods
- **Summary Statistics Using Parametric Methods**

Probability Plots

Probability plots are a method to check distributional assumptions. They give a visual check of of how well data conform to an assumed distribution. For example, we can check to see if data appear to be from a normal distribution (bell curve). If the probability plot shows the points falling roughly on a straight line, then the assumed distribution is appropriate for the data.

Normal Probability Plot (with censor data)



Normal Probability Plotting

The probability plot on the previous page was made based on water samples taken from the Thompson River near Savona. This data set can be found in the Appendices of the guidance document. Assuming this data has been read into an R workspace, and called `savona`, the following functions can be used to draw the probability plot seen on the previous slide.

```
>SavRos=with(savona,ros(obs=Orthophos,
 censored=OthoCen,forwardT=NULL))
>plot(SavRos)
```

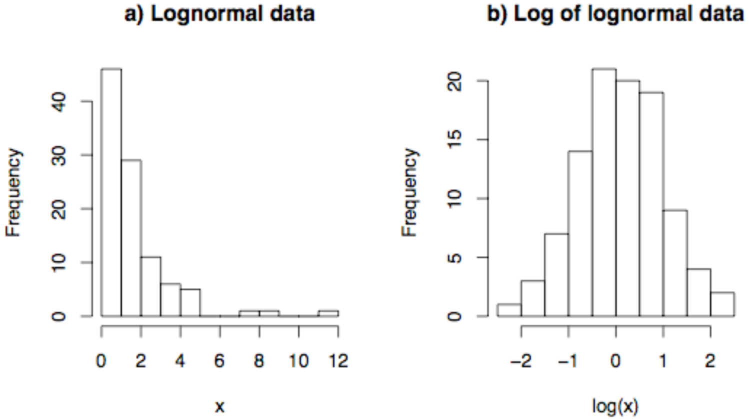
Notice the arguments to the function...

Lognormal Probability Plotting

The normal distribution is probably the best known distribution for data analysis, because it is commonly taught in first year statistics classes. When analyzing water quality data, there is another distribution that will frequently represent the data well; this is the *log-normal* distribution. To see if data from a sample appear to conform to a log-normal distribution, we can create a log-normal probability plot.

```
>SavRos2=with(savona,ros(obs=Orthophos,
 censored=OthoCen,forwardT="log"))
>plot(SavRos2)
```

Relationship Between Normal and Lognormal Distributions



Relationship Between Normal and Lognormal Distributions

- If we take the log of values from a log-normal distribution, the *transformed* values will be normally distributed
- We can *back-transform* values that have been log transformed using *arithmetic* expressions. The formulas for this can be found in the guidance document
- The *mean* of the log-transformed values will be the *median* of the values on the original scale
- Most of the data found in water quality (and similar) data sets will appear to follow either a normal or a lognormal distribution
- In the rare cases where data does not follow one of these two formats, you may have to consider alternative types of analyses. These are also discussed in the guidance document.

Statistics: A Quick Review

Statistics is a word that is generally defined in two ways. Understanding both of these definitions is useful in conducting good statistical analyses.

The first definition refers to statistics as a scientific discipline.

The second definition refers to the calculations we make based on sample data that we have observed.

Statistics

As a scientific discipline, statistics is concerned with converting *data* into *information* in order to facilitate decision making. As such, drawing graphs to summarize data are as much a part of statistics as making fancy numeric calculations.

Statistics is also often called the *science of uncertainty*

Statistics

Statistics can also be calculations that we make on a data sample. Understanding these values is also a key part of the process of turning data into information.

Statistics:Calculations

In statistics (the discipline) lingo, we have things called *Populations*, and characteristics we are interested in called *Parameters*. For example, we might be interested in the concentration of phthalates, a component of landfill leachate, in groundwater near a landfill.

In this example, our population is ALL of the groundwater near the landfill. The parameter we are interested in is the concentration of phthalates.

Statistics:Calculations

For obvious reasons, it is not practical to sample ALL of the groundwater near a landfill. Instead, the sensible approach is to take *Samples* of the groundwater, and calculate *Statistics* based on the sample.

We calculate *Sample Statistics* to make inferences about *Population Parameters*!

Statistics: Calculations

When writing up an analysis, statisticians generally use *greek* letters to represent the population parameter of interest. *Latin* letters are used to represent the corresponding sample statistics. For example

$$\mu \longrightarrow \bar{x} \text{ (mean/average)}$$
$$\sigma \longrightarrow s \text{ (standard deviation)}$$

Analysis Approaches

There are five basic approaches to approaching data that has censored values where the percent of data that is censored is less than 50%

- Deletion (very, very bad!!!)
- Substitution (very, very bad!!!)
- Parametric Methods
- Non-Parametric Methods
- Semi-Parametric Methods

Deletion

Just because you can't see it, does that mean it isn't there?

Doing this is a good way to blow up a spaceship!!

Or at least leave yourself open to litigation.

Substitution

As bad as deletion! Common values for substitution are 0, 1/2 the censor limit, and the censor limit itself. We saw in the boxplot example what erratic results this can give!

There is no theoretical OR practical justification for doing this. It does not give consistent results, and is therefore not scientific!

This could also leave you liable to litigation. There are well publicized, much better alternatives available!!

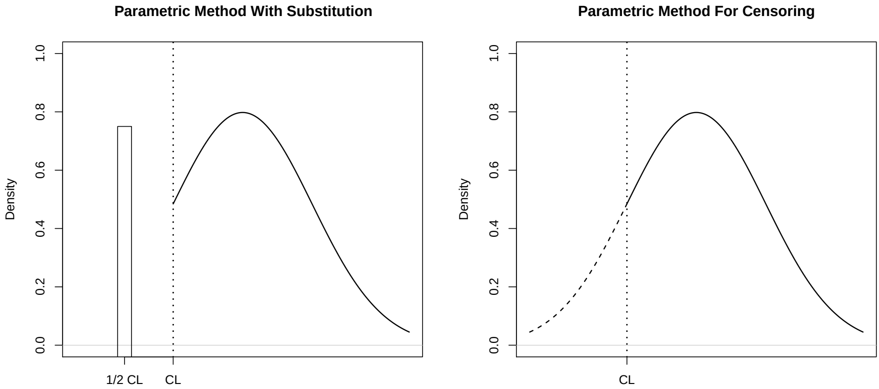
Much, much better alternatives!!

Parametric Methods

In parametric methods, we assume that the data follow some known distribution, such as the normal or lognormal distributions. This assumption includes data both above and below the censor limit. Estimates are made using maximum likelihood methods, where the distribution of values below the censor limit is imputed based on the distribution of values above the censor limit.

For this method to be accurate we need to assess whether our distributional assumptions are realistic. (Think plots)

Substitution vs. Parametric Methods



Non-Parametric Methods

When the distributional assumptions for parametric methods are not met, non-parametric methods are a good alternative analysis method. Non-parametric methods focus on the *ordering* of the data rather than on its distributional form. For example, using these methods the measure of centre becomes the median (50^{th} percentile) rather than the mean.

The assumptions for this method are fewer, and it handles very extreme values better than parametric methods.

Semi-Parametric Methods

This analysis method combines some of the best features of parametric estimation with some of the best features of non-parametric estimation.

Semi-parametric methods are probably the optimal method for obtaining summary statistics.

Unfortunately, these methods have not been extended for use in more complicated analyses, and so are limited in scope.

Summary Statistics

Although all of the statistical methods introduced above are shown in the guidance document, for the rest of this course I am going to focus on parametric methods using maximum likelihood estimation.

These methods are quite similar in principal to the standard statistical methods that are taught in statistics classes, such as z and t-tests.

Although there is technically no minimum sample size when using this method, its use is generally restricted to samples large enough that distributional assumptions can be checked using boxplots and normal probability plots.

Summary Statistics

When calculating summary statistics it can be useful to conduct the analysis in the following order

1. Simple Data Summaries
2. Plotting
3. Summary Statistics

Conducting analyses in this order minimizes unnecessary data analysis steps by assessing data assumptions as you go.

Simple Data Summaries

The first step prior to conducting an analysis is to determine what the percentage of censoring is in your data. The NADA package in R has a function to do this for you that I will demonstrate using the savona data discussed earlier

```
>censummary(obs=savona$obs,censored=savona$cen)

all:
      n   n.cen pct.cen     min     max
 32.000   7.000  21.875   0.001   0.009

limits:
  limit n uncen pexceed
1 0.001 7    25 0.78125
```

Simple Data Summaries

As long as the summary of the data indicates that the percent censoring is less than 50%, it is okay to go ahead with any of the analysis methods listed.

If more than 50% of the data are censored, it is still possible to use non-parametric methods to summarize the data that is greater than the median.

When the data was to be used for analysis using covariates, *generalized linear models* can be used. These are discussed in the last chapter of the guidance document.

Plotting

Subsequent to looking at the summary of the percentage censoring that is present in the data, it is useful to assess model assumptions for parametric methods using plots. The main model assumptions for creating data summaries using parametric methods are *independence* (true of all the methods described in the guidance document), and the normal or lognormal distributional assumptions.

We have already described how to draw and interpret plots on earlier slides.

Summary Statistics

Following an evaluation of whether the percent censoring leaves the data suitable for analysis, and whether a normal or lognormal distribution assumption are appropriate, we can generate summary statistics on data using parametric methods.

If censoring is high, or we cannot find a suitable distribution to represent the data, it would be necessary to consider other data analysis alternatives.

Summary Statistics

There are several steps to creating a complete set of summary statistics using parametric methods in R. The first is to created a *censored MLE* object, which can then be used to provide the desired estimates such as the mean, standard deviation, and quantiles for the distribution as estimated from the data.

This is done using the `cenmle()` function in R.

Summary Statistics

The syntax for using this function is demonstrated using the Savona data assuming a *normal* distribution.

```
>SavMLE=with(savona,cenmle(obs=Orthophos,
censored=OthoCen,dist=‘gaussian’))
```

Notice the argument `dist=’gaussian’`, this is how we specify that the distribution to be used is the *normal*, or *gaussian* distribution.

If we wanted to conduct estimates assuming a lognormal distribution we would specify the argument as `dist=‘lognormal’`

Summary Statistics

Using the function shown above results in the following output in R

```
>SavMLE

           n        n.cen       median
32.000000000  7.000000000  0.003492687
        mean           sd
 0.003492687  0.002550961
```

Summary Statistics

Mean, standard deviation, and quantile estimates can also be obtained from the `cenmle` object that we have created. These values are also reported with appropriate 95% confidence intervals to help express our uncertainty in the estimates.

```
 >mean(SavMLE)

        mean           se      0.95LCL
0.0034926866 0.0004515784 0.0026076092
     0.95UCL
0.0043777639
```

Summary Statistics

```
>sd(SavMLE)

[1] 0.002550961
```

Summary Statistics

```
>quantile(SavMLE,conf.int=TRUE)

  quantile         value       0.95LCL
1     0.05 -0.0007032712 -0.0024394335
2     0.10  0.0002234983 -0.0013289567
3     0.25  0.0017720894  0.0005223881
4     0.50  0.0034926866  0.0025729670
5     0.75  0.0052132837  0.0039635825
6     0.90  0.0067618749  0.0052094199
7     0.95  0.0076886443  0.0059524820
      0.95UCL
1 0.001032891
2 0.001775953
3 0.003021791
4 0.004412406
5 0.006462985
6 0.008314330
7 0.009424807
```

Summary Statistics

Having now obtained all the information and graphs to validate that you are conducting an appropriate analysis, we now also have the output of such an analysis, and the computer commands that were used to create it. Having all of these components is useful both when writing your own reports, or when reviewing the analysis/reports of other individuals!!

The End!!!!

Thank You!