**1 – Univariate Analysis of the Response Variable ()**

In this section we will be reviewing statistical methods for analyzing a single numeric variable (). Later we will examine the relationship between and a set of other variables or predictors (’s).

Topics reviewed in this section:

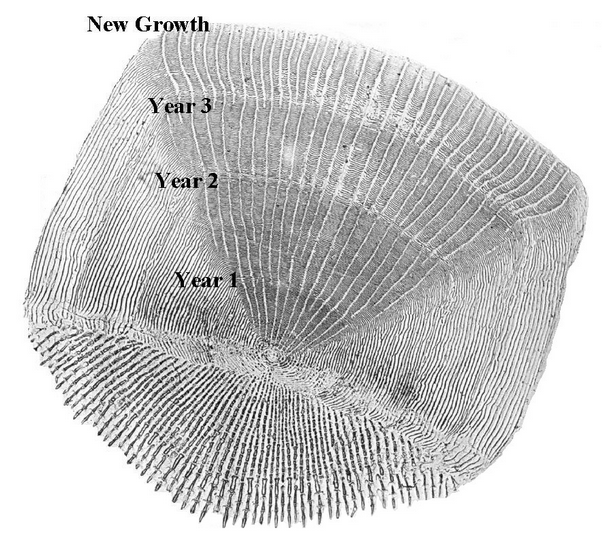
* Descriptive analyses for a single numeric variable   
  (numeric & visual summaries)
* Inferential methods for a single numeric variable

(hypothesis/significance testing, confidence intervals, prediction intervals)

* Connection to regression analysis (not review)

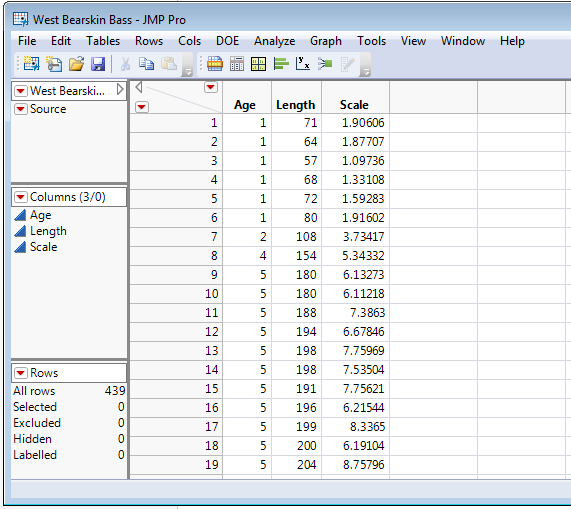
**Example 1.1 – Lengths of Smallmouth Bass in West Bearskin Lake**

A sample of *n = 439* smallmouth bass was taken by the Minnesota Dept. of Natural Resources (DNR) from West Bearskin Lake in northern MN and their length, age, and scale radius were measured. The age of smallmouth bass can be determined by counting growth rings on their scales and thus is also related to scale radius. Suppose researchers are interested in determining if the mean length

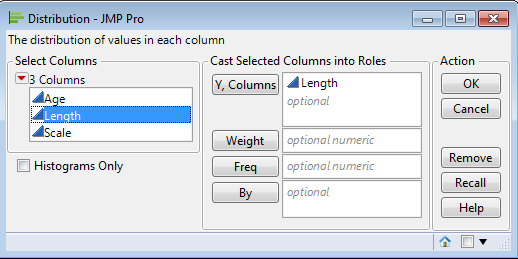
If view length (mm) as the response of interest, i.e. = length (mm) let’s consider basic descriptive statistics for the bass sampled.   
  
**Research Question:** Suppose researchers are interested in determining if the mean length of smallmouth bass in this population is less than 8 inches (203.2 mm).

**Data in JMP**

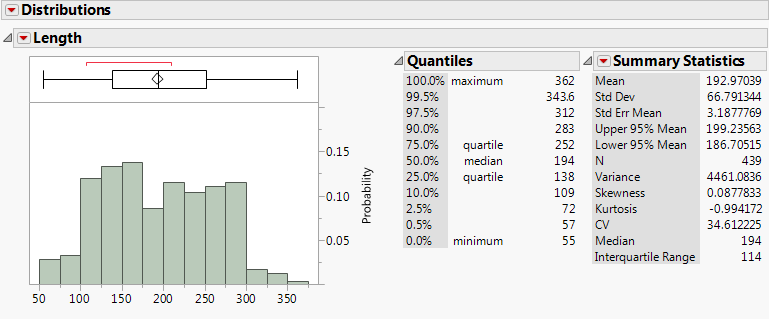


**1.1 – Descriptive or Summary Statistics for a Numeric Variable**

To obtain univariate descriptive statistics for the length of the smallmouth bass select **Analyze > Distribution** and double-click Length to move it to the list of variables on the right.



The somewhat customized result is shown below:



Here I have selected the following additional summary statistics from the **Summary Statistics > Customize Summary Statistics** option: sample variance, skewness, kurtosis, coefficient of variation (CV), median, and interquartile range (IQR).

**Discussion:**

Sample mean )

Sample variance ()

Sample standard deviation ()

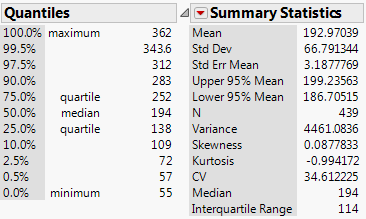
Coefficient of Variation (CV)

Skewness

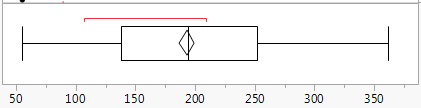
Kurtosis

Median

Interquartile Range (IQR)



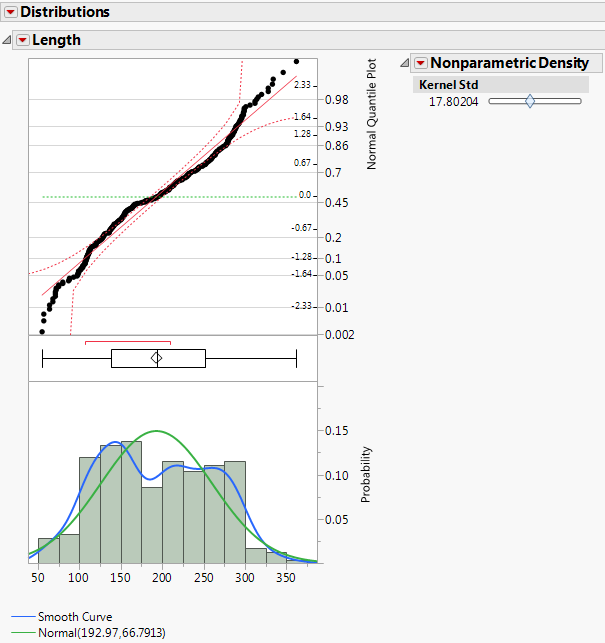
Outlier Boxplot for Smallmouth Bass Length (mm)

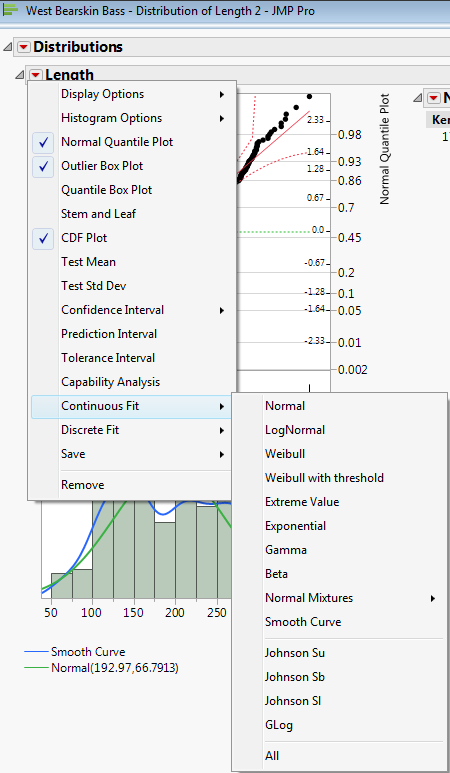


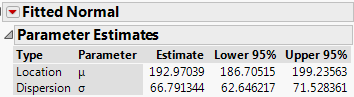
**1.2 – Univariate Displays or Graphs for a Numeric Variable**

For examining the distribution of a single numeric variable we typically use plots such as: histograms, dotplots, rug plots, boxplots, symmetry plots, density estimates, distributional fits, CDF plots, and normal quantile plots. There are others but we will use all of these at times in this course.

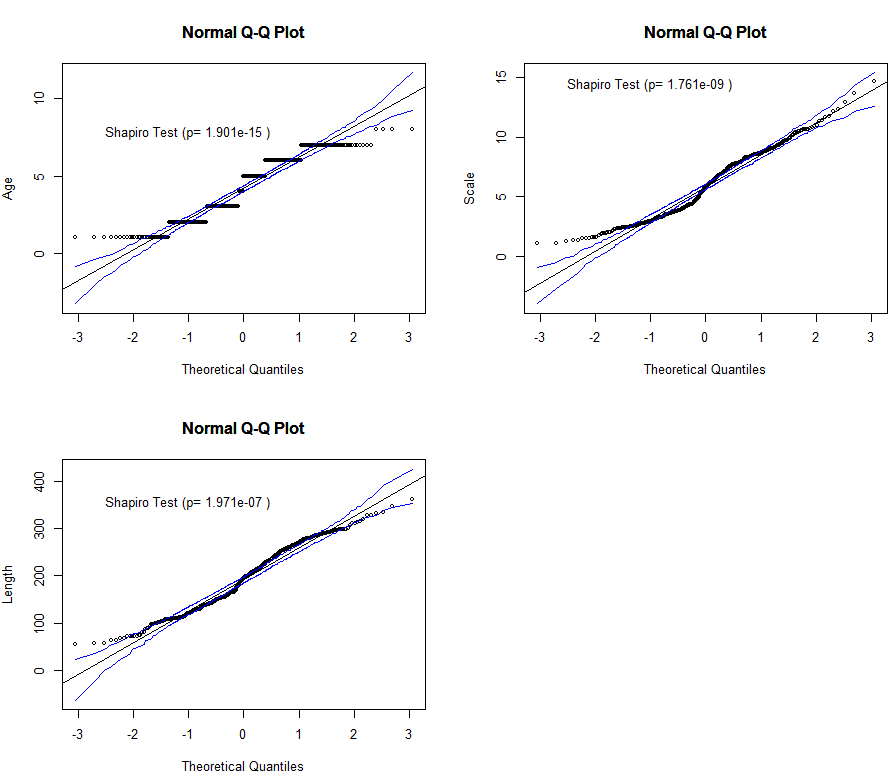
Below are some common visual tools for examining the distribution of a numeric response.



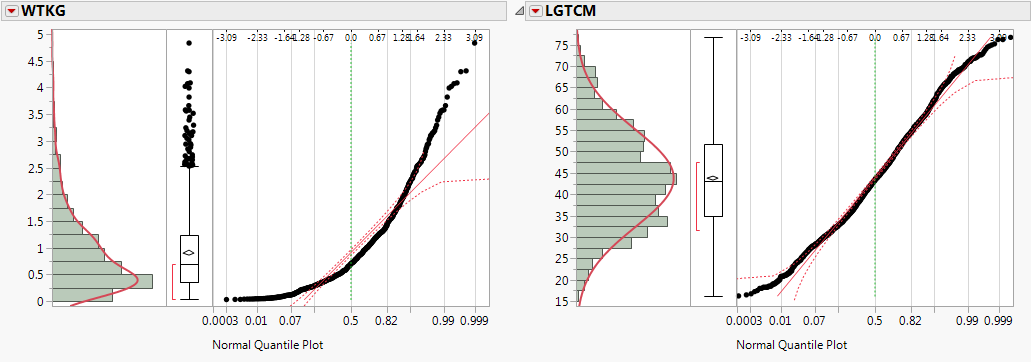


  
  
Notes:

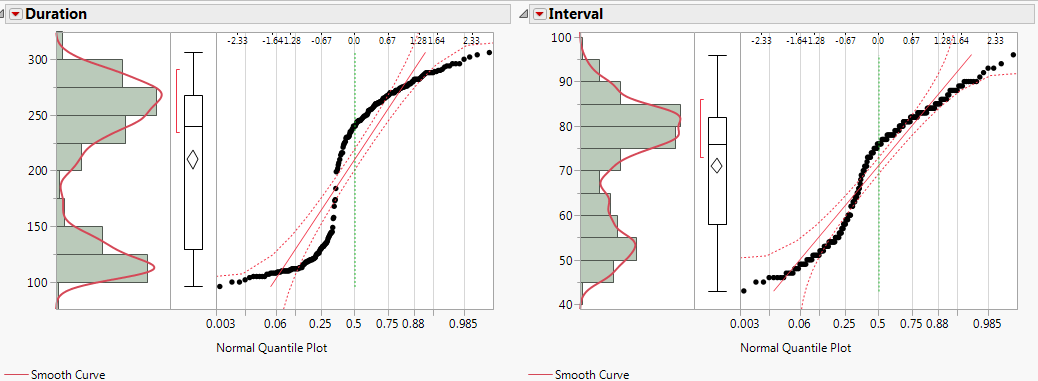
**Normal Quantile Plots – Smallmouth Bass Study**



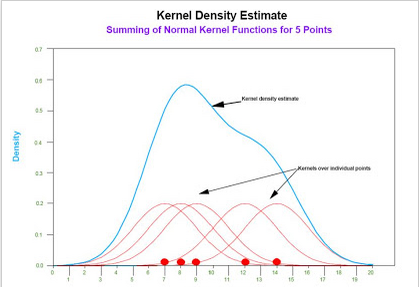
**MN Walleyes Study – from a study of mercury contamination in walleyes**

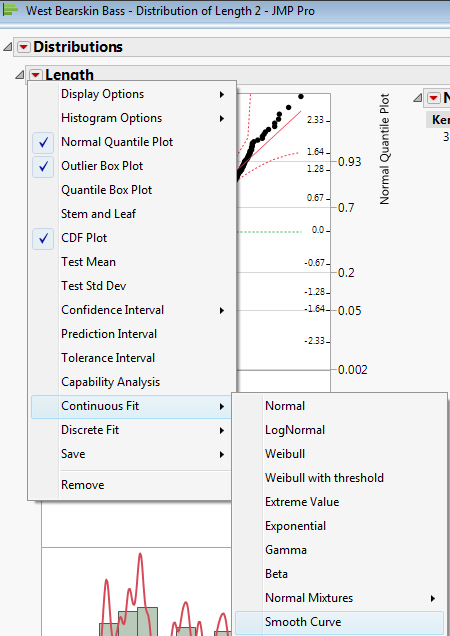


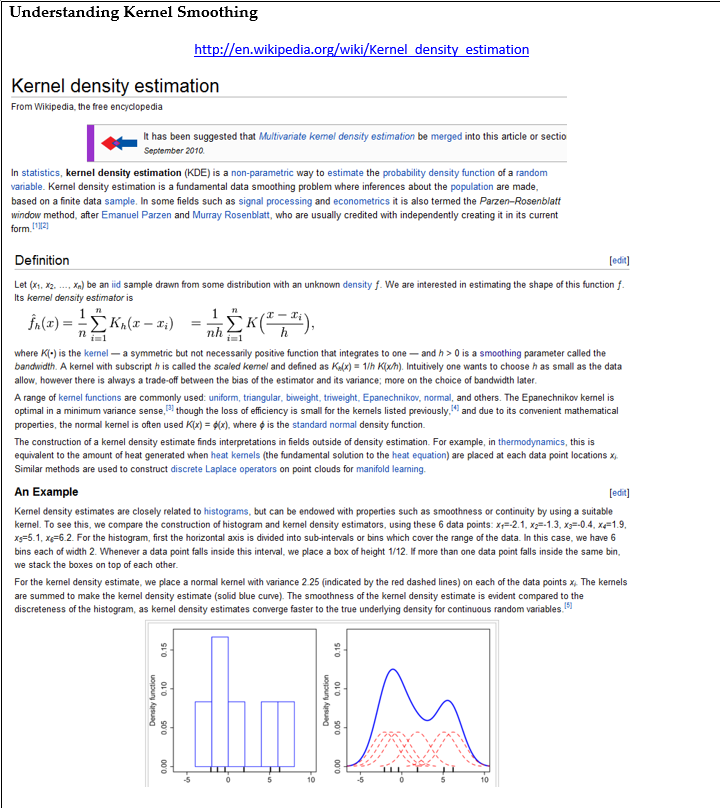
**Old Faithful Geyser Eruption Lengths** (Duration) **& Waiting Times** (Interval)



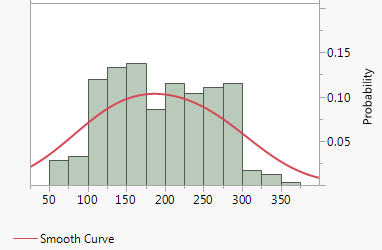
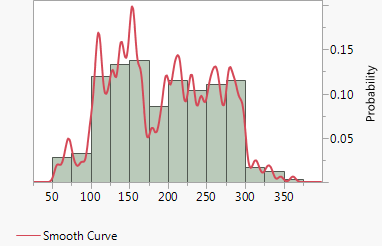
**Kernel Density Estimates** (Details of the Continuous Fit > Smooth Curve fit in JMP)







KDE for Length - Too smooth KDE for Length - Too noisy

Histogram, KDE, and Rug Plot in R  
> data(wblake)

> attach(wblake)

> names(wblake)

[1] "Age" "Length" "Scale"   
> head(wblake)  
> head(wblake)

Age Length Scale

1 1 71 1.90606

2 1 64 1.87707

3 1 57 1.09736

4 1 68 1.33108

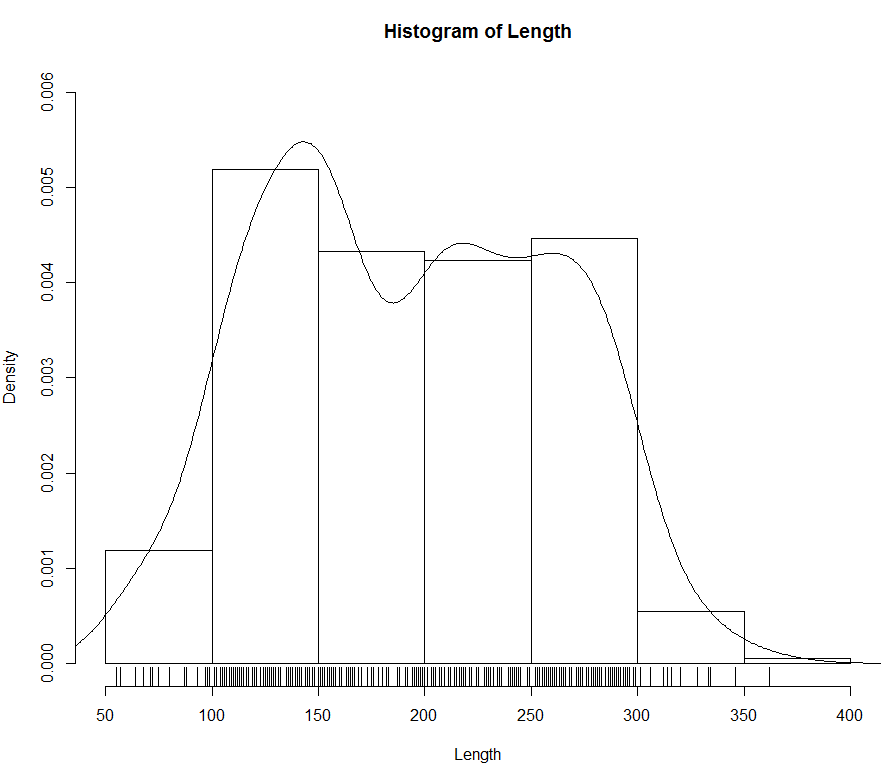
5 1 72 1.59283

6 1 80 1.91602

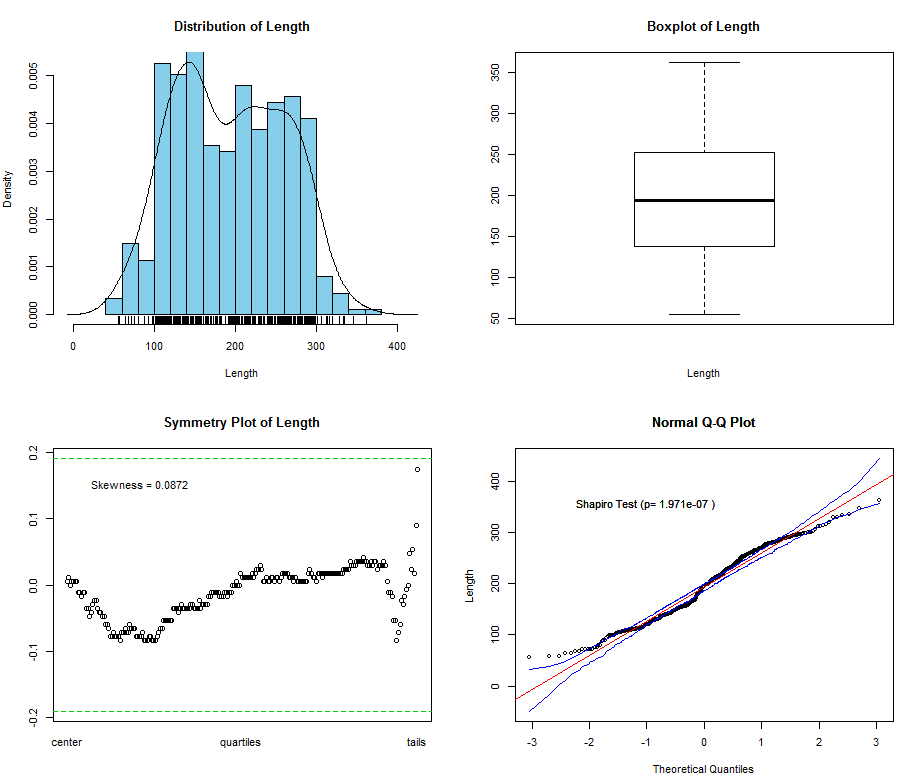
> hist(Length,prob=T,ylim=c(0,.006))

> lines(density(Length))

> rug(Length)

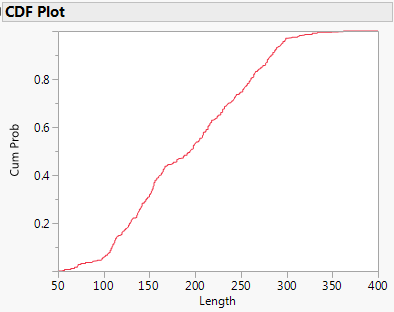


In the mult.RData directory you will find a function I wrote called Statplot which gives four univariate displays for a numeric variable: a histogram with KDE density estimate and rug plot, a boxplot, a symmetry plot, and a normal quantile plot with a *simulated envelope*.

> Statplot(Length)  


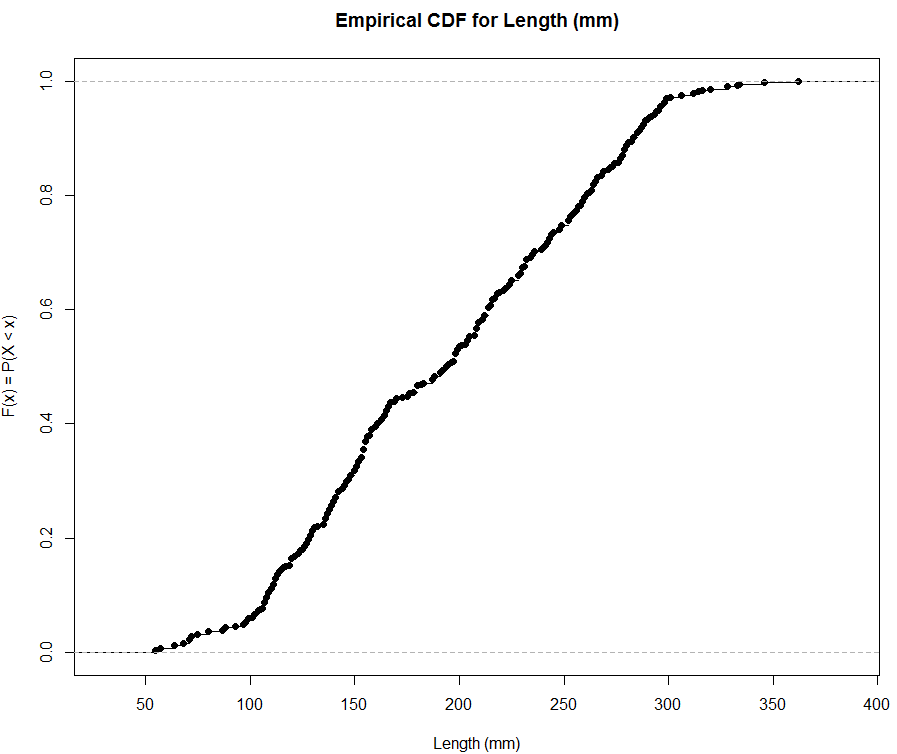
**Notes:**

Empirical Cumulative Distribution Function



Empirical CDF in R

> plot(ecdf(Length),xlab="Length (mm)",  
ylab="F(x) = P(X < x)",main="Empirical CDF for Length (mm)")



**Notes:**

**1.3 - Univariate Transformations (Ladder of Powers)**

As a general rule or guideline, normality is your friend when working with continuous numeric data and this is certainly true in regression. In reality, no continuous variable can have a distribution that is theoretically normally distributed, as the normal distribution is defined over the entire real line (i.e.

One of the most common departures from normality that we see when working with continuous numeric data is skewness – either left or more commonly right.

**Ladder of Powers (Tukey, 1977)**

Scaled Power Transformation Family

Tukey Power Transformation Family

**Example 1.2: Mercury Contamination in MN Walleyes**

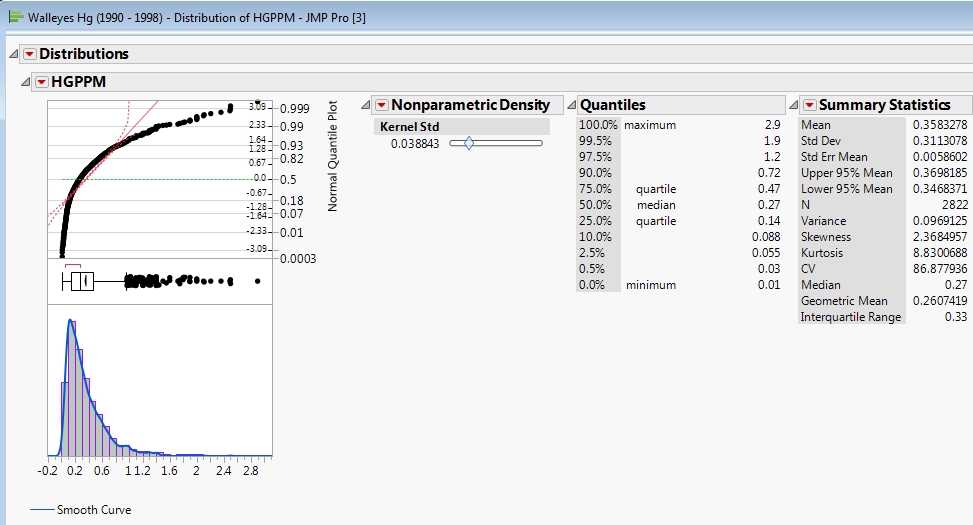
These data come from Minnesota's Fish Contaminant Monitoring Program (FCMP) which is a joint effort by the DNR, Minnesota Dept. of Health, MN Dept. of Agriculture (MDA), and the Minnesota Pollution Control Agency (MPCA) from the years (1990 -1998).

**Data Files:** **Walleyes Hg (1990-1998).JMP** and **Walleyes Hg (1990-1998) Major Waterways.JMP** - the first file contains the results from all waterways sampled in this time period and the second contains data from the most sampled waterways in the first database. I have created datasets for two waterways near Duluth, MN - Island and Fish Lake Flowage only (**Walleyes Fish vs. Island.JMP**, **Island Lake.JMP**, **Fish Lake Flowage.JMP**)   
  
**Data Description:**

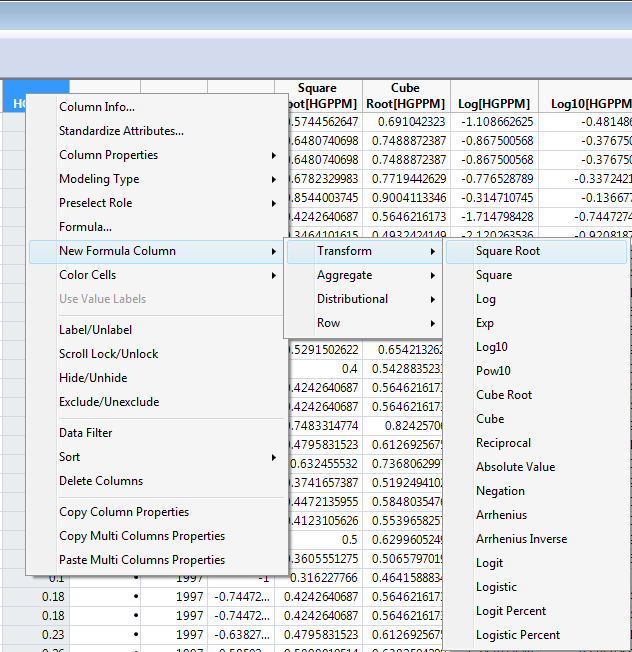
* WATERWAY – waterway where walleyes were sampled
* LOCATION – sampling location within waterway
* CTY – county number
* DATECOL – date collected
* NOFISH – number of fish used (typically 1, but they did combine fillet samples in some cases)
* LGTHIN – length of fish (in.)
* WTLB – weight of fish (lbs.)
* LGTCM – length of fish (cm)
* WTKG – weight of fish (kg)
* Latitude – latitude of sampling location (mostly missing)
* Longitude – longitude of sampling location (mostly missing)
* HGPPM – mercury level (ppm)
* PCBPPM – PCB level (ppb – mostly missing)
* Year – study year (1990-1998)
* log(Hg) – log base 10 of mercury level (log10(ppm))

We will be using these data later in the course to develop walleye consumption advisories using regression analysis. For now we will focus on the distribution of the response variable HGPPM which contains the mercury concentration in parts per million (ppm) of the walleyes sampled.

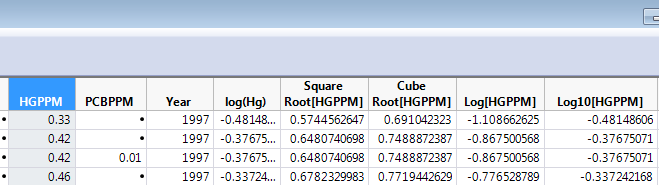
Below is univariate summary of HGPPM in JMP with all the “bells & whistles”.



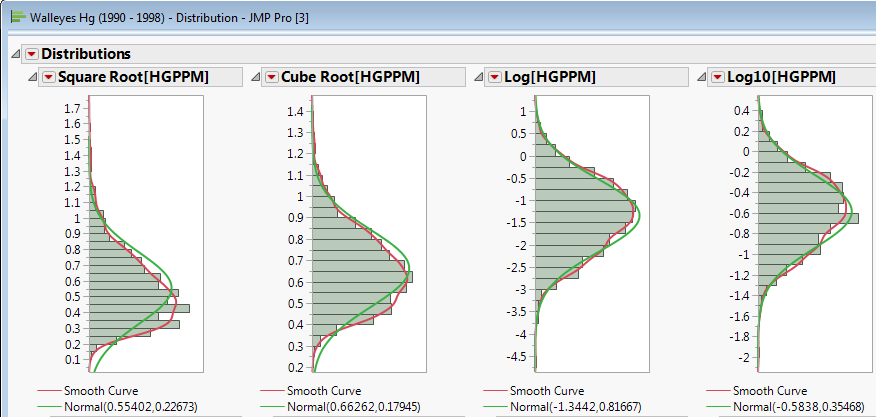
Clearly the distribution of the mercury contamination found in the n = 2822 sampled walleyes is markedly right or positively skewed. We can use Tukey’s Ladder of Powers to find a power transformation of mercury level that is approximately normally distributed. Easily power transform a variable in JMP right-click at the top of the column for the variable you wish to transform and select **New Column Formula > Transform** as shown below.



The results of the square root, cube root, natural logarithm (base e), and logarithm (base 10) are shown below.



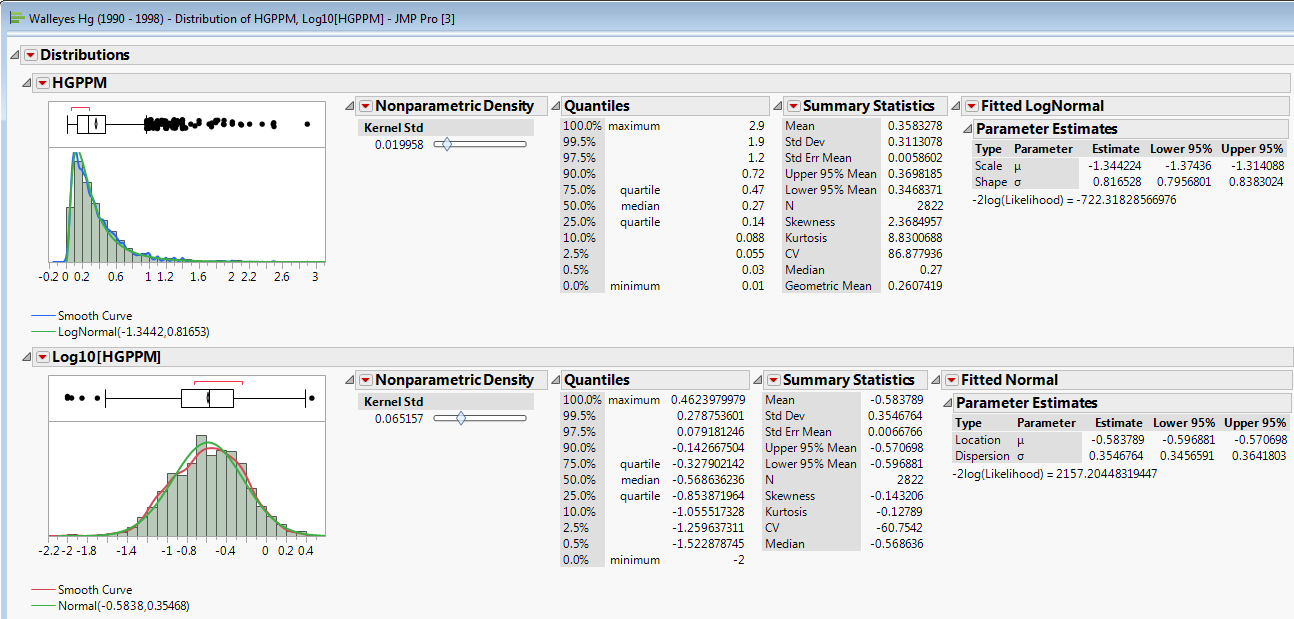
Histograms of with KDE and Normal density estimates added are shown below.



**Notes:**

The log transformation is the definitely the most common transformation used in statistics. From the mercury concentration example above we can see that the log base used is irrelevant, as the distributional shape is identical for both bases used above (*e* and 10). Another common base to use, particularly in regression modeling, is base 2 which we will see in examples later in the course.

Below are summary statistics and graphical displays of



**Notes:**

**1.4 - Inferential Methods for a Single Numeric Variable**

In this section we will review inferential methods for a single numeric variable, namely hypothesis or significance tests and confidence intervals. I assume you are familiar with these from your prerequisite coursework, however the notation we will be using in this course will likely be different.

**Example 1.1 – Lengths of Smallmouth Bass in West Bearskin Lake (cont’d)**



**Research Question:** Researchers are interested in determining if the mean length of smallmouth bass in West Bearskin Lake is less than 8 inches or 203.2 mm.

Converting the research question into standard notation:

**Parameter of Interest**

**Statement of Hypotheses**

(pop. mean length is not less than 200 mm)

(pop. mean length is less than 200 mm)

For testing a single population mean we generally use a t-test. The use a t-test requires that either:

* the variable in question is approximately normally distributed

or if

* we have a “large” sample size (*n > 30* typically, but that is not set in stone)

If neither of these conditions apply the use of the standard t-test is questionable.

For these data the sample size is large (n = 439) so even though the distribution is probably not approximately normal (slightly bimodal) we can safely use a t-test and confidence interval for the population mean to answer the research question.

The t-test test statistic is given by:

**Notes:**

**Hypothesis Testing for a Single Population Mean ()**

|  |  |  |
| --- | --- | --- |
| Lower-Tailed Test  (AH contains <) | Upper-Tailed Test  (AHcontains >) | Two-Tailed Test  (AHcontains ≠) |
|  |  |  |

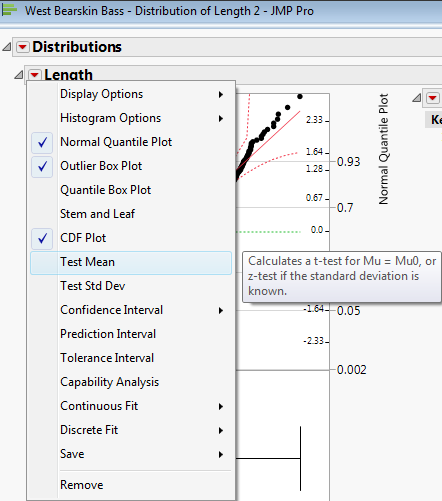
The *p-value* is the probability of obtaining a test statistic value (i.e. a t-statistic value) as extreme or more extreme than the one observed assuming the NH is true. Here the t-statistic is *t = -2.2052* which gives a lower-tail *p-value = .01397*.

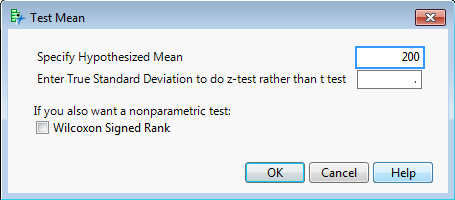
For **all** statistical tests we reject the null hypotheses (NH) if the p-value associated with our test statistic is less than the pre-chosen significance level (), which is typically .

Using R as a probability calculator:

> pt(-2.2052,df=438,lower.tail=T)  
[1] 0.01397947

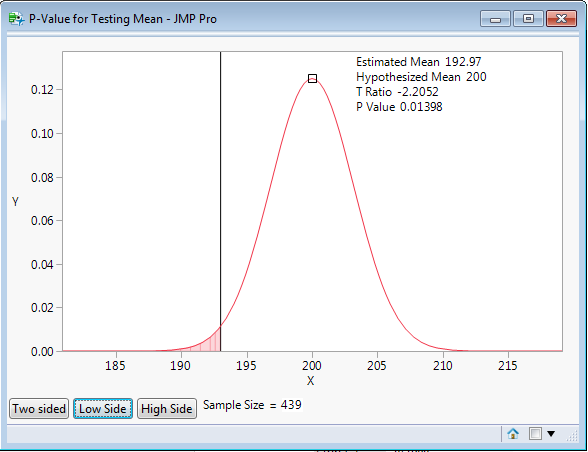
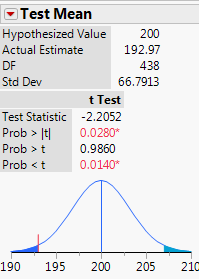
**t-Test in JMP**

Starting from the results of the Distribution for Length we select Test Mean from Length drop-down menu as shown below and then input 200 (mm) as the hypothesized mean in the t-Test dialog box.



NH:

AH:



**Conclusion from t-test:**

We have evidence to conclude that the mean length of smallmouth bass in West Bearskin Lake is less than 200 mm (p = .014).

Tests of significance really don’t tell us much if you think about it. It only says that we have evidence to suggest that , but it really doesn’t say what we think the mean actually is! That is where confidence intervals come in.

The basic form for most (but not all) confidence intervals is given by:

Or more explicitly

Margin of Error (*E*)

**100(1-)% CI for the Population Mean ()**

Assuming the use of the t-distribution is appropriate for the response of interest the CI for the population mean is given by:

Where is the quantile of the t-distribution with degrees of freedom (*df*).

**Notes:**

In R   
Assuming using (or 95% confidence

> qt(.025,438)

[1] -1.965395

Assuming using (or 99% confidence

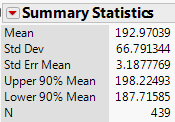
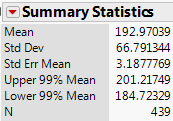
> qt(.005,438)

[1] -2.5871

Here a 95% CI for the mean length of smallmouth bass in West Bearskin Lake is given by:

192.97 =

Anytime you use **Analyze > Distribution** in JMP you obtain a 95% CI for the population mean (). You can change the confidence level in the **Customize Summary Statistics** menu.  
  
95% CI for E(Length) 99% CI for E(Length) 90% CI for E(Length)



**Notes:**

**Prediction Intervals** (not review)

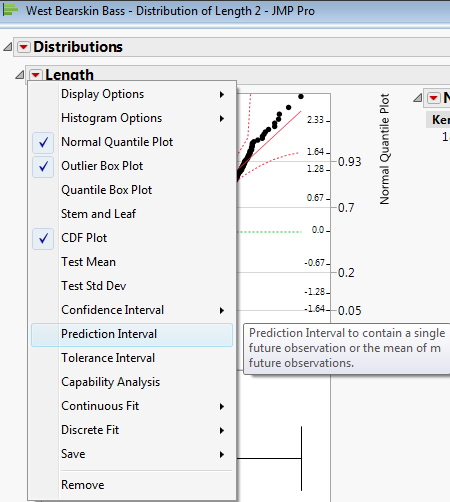
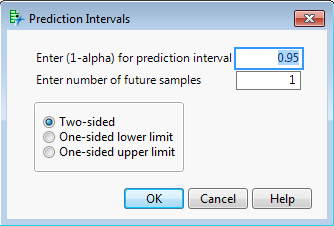
What if we wanted a range of lengths that I am 95% confident will cover the length of a single fish sampled (e.g. caught on my fly rod). The interval will have to wider because I also have to capture the fish-to-fish variability in addition to my uncertainty about the population mean.

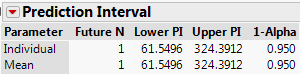
For predicting a single new value (or fish length in this case) the standard error formula is given by,

SE( or sepred =

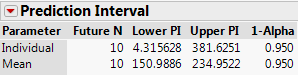
Thus a 95% prediction interval for the length of a single fish sampled from the lake is:

**Prediction Intervals in JMP**



For predicting the sample mean of *k* future fish sampled (e.g. k = 10).



**1.5 – Inference for the Population Mean Viewed as a Regression Model**

We will now examine inference for a single population mean viewed as a regression model. When developing a regression model we first specify **mean function** and a **variance function**. We can then also formulate the model in terms a random sample of size *n* from the population. If we are willing to assume the lengths of smallmouth bass in W. Bearskin Lake are normally distributed with mean and variance , i.e. , then we have the following.

Mean Function

Variance Function

Data-based Model

The data-based model represents the actual values of Y we observe based on sample of size *n*.

* Random sample:
* Model:

🡨 This will look more familiar

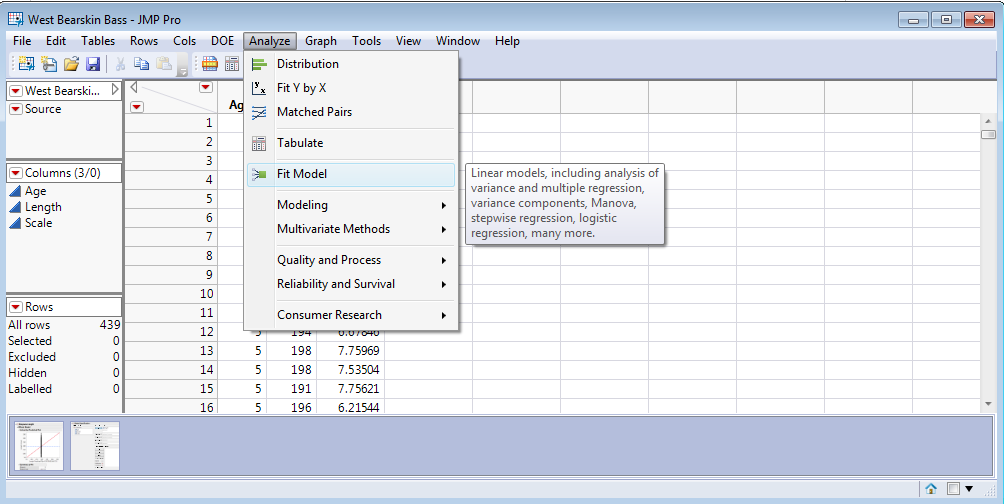
later. This is the intercept   
 only regression model.

This model says that when we observe the length of the fish sampled we observe mean plus random variation about the mean which is captured by the error term . Note this does not represent measurement error, but rather the fact that all fish are not all the same length!

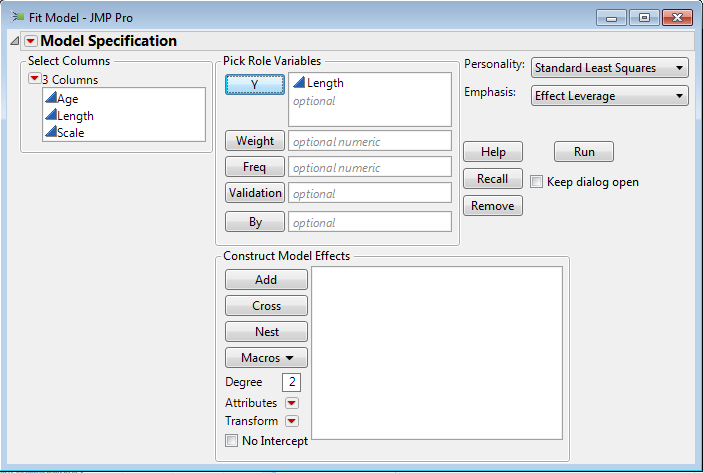
If we assume that then we have that the errors are also normally distributed, i.e. .

**Fitting our model in JMP**

To fit this simple model in JMP, select **Analyze > Fit Model** then set up the Fit Model dialog box as shown below.



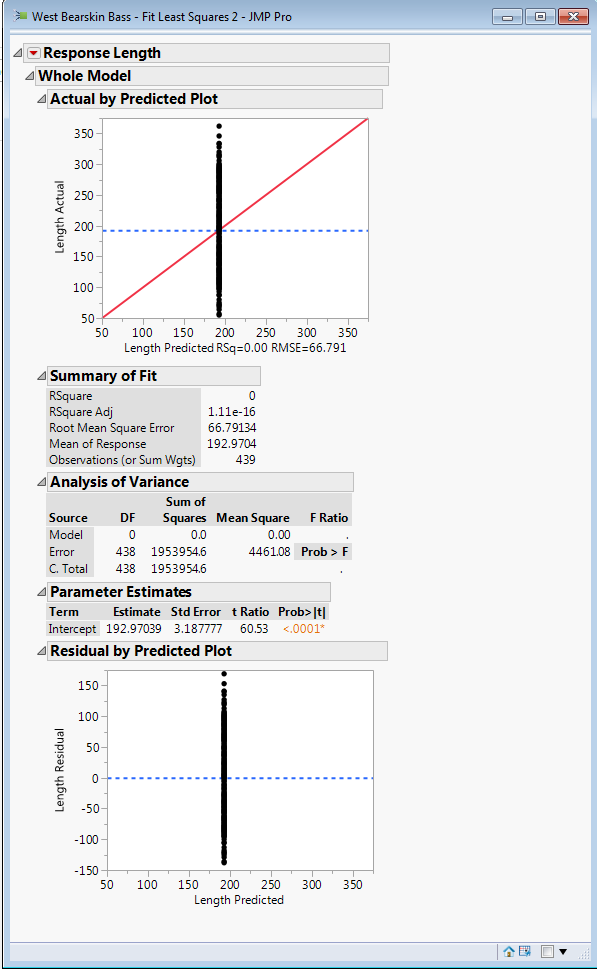
**Fit Model** dialog box



The results of fitting this model are shown on the following page along with some of our results from examining the distribution of the fish lengths.

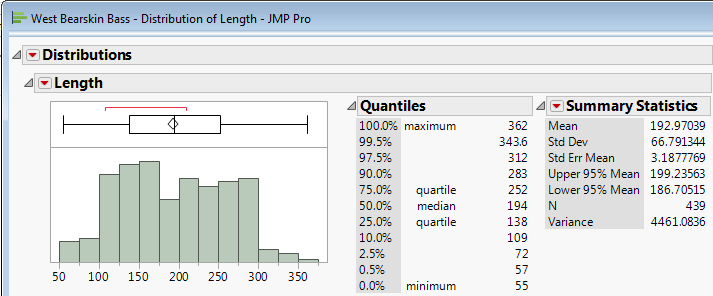
This will fit the equivalent regression models defined above, i.e.

Results of **Fit Model**



**How are the results similar?**

Results from **Analyze > Distribution**



**The Fitted Model**

Models

Fitted Models

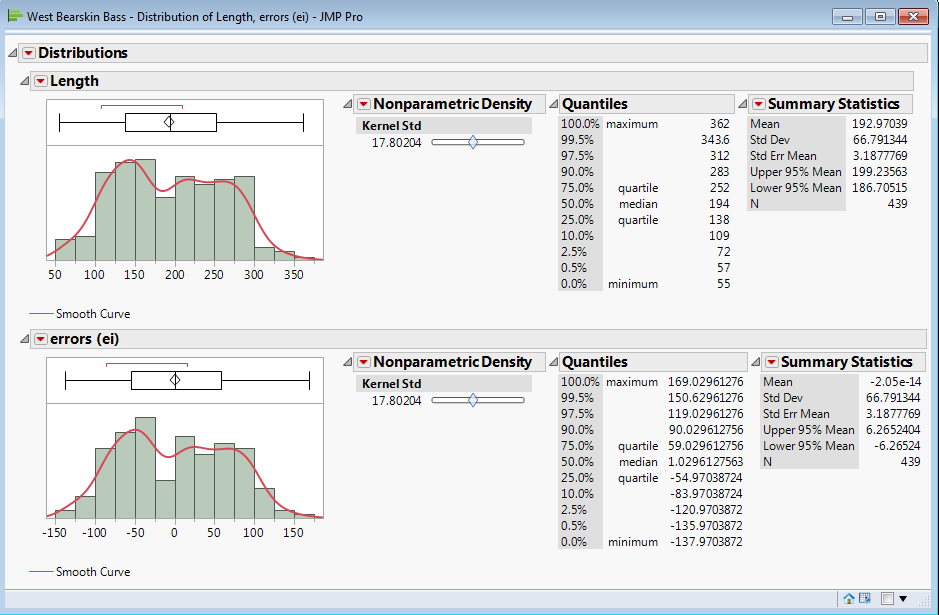
where,

-- thus the ***fitted values*** are

and the observed errors or ***residuals*** ( are given by

.

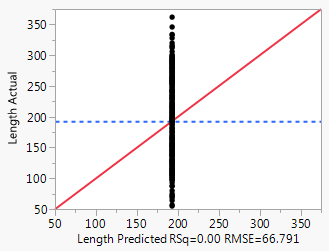
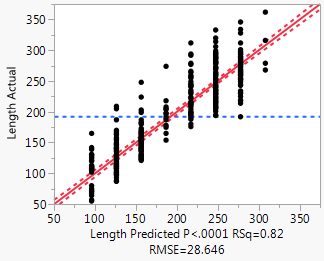
**Distribution of Observed Lengths ( and the Residuals ()**

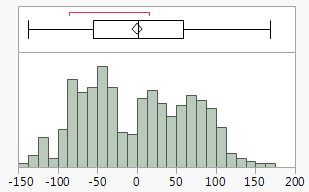
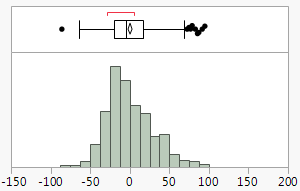


**Notes:**

**1.6 – Briefly Looking Ahead**

What do we gain by using information about the age (yrs.) of the smallmouth bass in modeling the length?

RSS = 1953954.6 RSS = 358595.5 🡨 81% reduction!