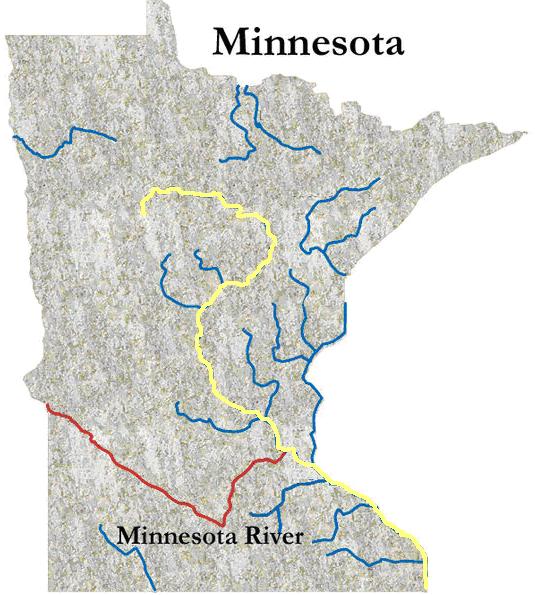
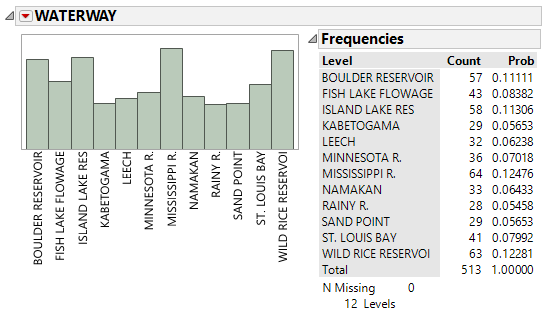
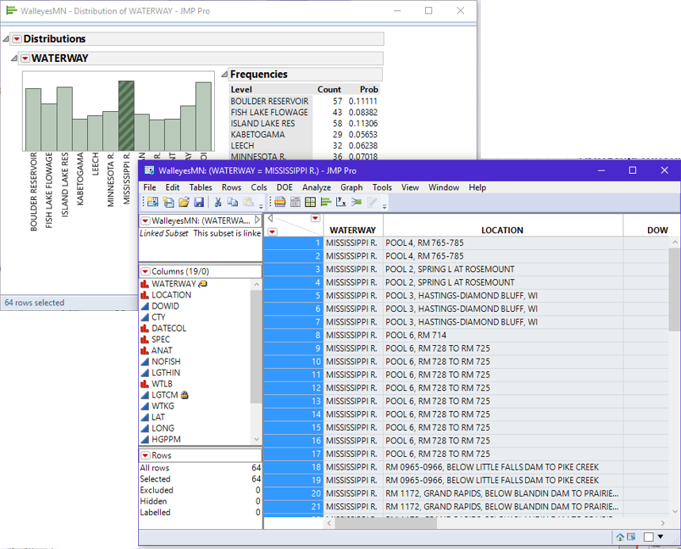
**Summary Measures for Numerical Data**

**Example 9.1** Consider an investigation of mercury levels in Walleyes along the Mississippi River in MN. Open the WalleyesMN file provided on the course website.

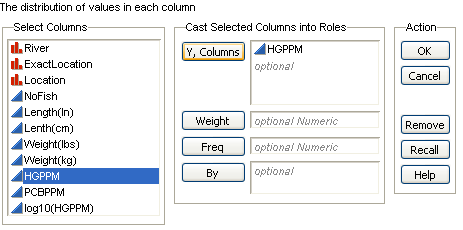




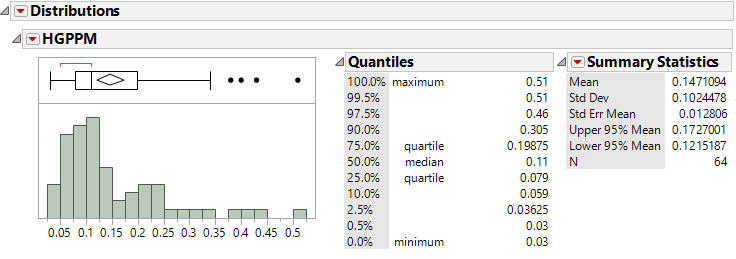
First, we will consider only the observations from the Mississippi River. This will allow us to investigate the HGPPM levels of Walleye along the entire Mississippi River.



Once the observations from the Mississippi River have been selected out, select Analyze > Distribution. Place HGPPM in the Y, Columns box.



The output from JMP…



Looking at each piece of the output…

|  |  |
| --- | --- |
|  | A |
| B |
| C |
| D / E |
| F |

|  |  |
| --- | --- |
|  | A |
| B |

|  |  |
| --- | --- |
| Histogram |  |
| Comment: A well-known problem with histogram is the number of bins used to make the histogram may change your perception of shape for the data. | |
|  |  |

|  |
| --- |
| Boxplot |

Identifying Outliers on Boxplots…

|  |  |  |
| --- | --- | --- |
| |  |  | | --- | --- | |  | There are four outliers in this dataset. Identify the location of each outlier. | |

Density Smoothers…

|  |  |
| --- | --- |
|  | To adjust the amount of smoothing… |

Normal Quantile Plot…

|  |  |
| --- | --- |
|  |  |

Transformations – An approach to fix issues with associated with skewness in data…

|  |  |
| --- | --- |
| Summary statistics on the log10 scale | Distribution on the log10 scale |

Cumulative Density Plots…

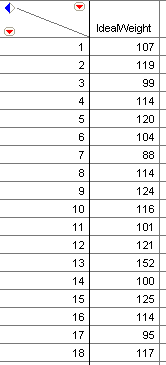
|  |  |
| --- | --- |
|  |  |
| CDF’s are a powerful way to compare two (or more) distributions. This plot was constructed using Analyze > Fit Y by X. Place HGPPM in the Y, Response box and Waterway in the X, Factor box. From the red drop down arrow, select CDF plot. Lastly, use a local data filter to select only the Minnesota and Mississippi Rivers to display. | |

Z-Scores – Rules for Outliers…

|  |  |
| --- | --- |
| Getting Z-Scores in JMP… | Outlier Rule:  z_score_scale |
| How are Z-Scores computed? | |
|  | |

**Inferential Methods for Numerical Data – Single Variable**

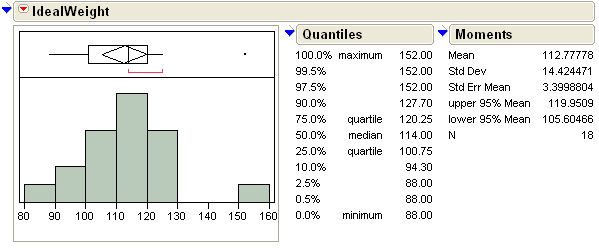
**Example 9.2** Consider a study where the weight of insulin-dependent diabetics is being investigated. The variable of interest is the percent of their ideal body weight. For example, a value of 120% implies that the individual weights 20% more than their ideal weight and a value of 95% implies the individual weights 5% less than their ideal weight.



Getting the summary measurements for the differences in JMP

|  |  |
| --- | --- |
|  |  |

Summary Statistics for this data…



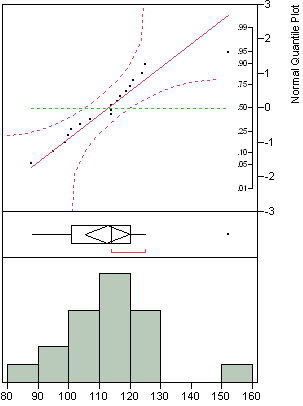
A value of 100 means an individual is at their ideal body weight. Discuss the summary measurements of these insulin-dependent diabetics.

Before going ahead with the inferential procedures, we should consider the following assumption for our testing procedure.

**Assumptions:**

|  |  |
| --- | --- |
| j0213523[1] | The observations from must follow a normal distribution |

This assumption can be checked in JMP using a normal quantile plot. A normal quantile plot can be obtained in JMP by selecting Normal Quantile Plot from the drop down menu. Another approach to verifying the normality assumption is to overlay a normal curve on top of the histogram. Both are shown next.



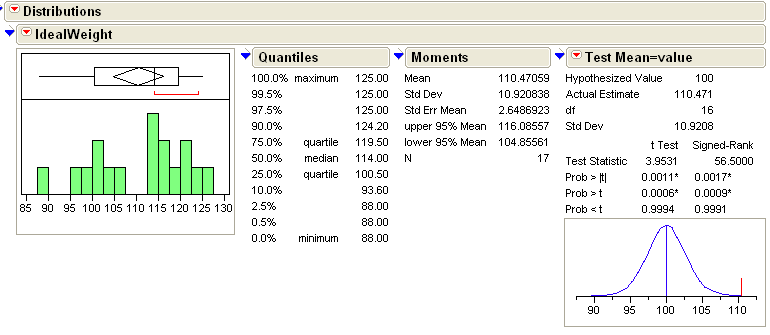
Question: Is the normality assumption satisfied? Discuss.

|  |  |
| --- | --- |
| Step 0 | **Research Question:** Is there statistical evidence to suggest that the average percent ideal body weight for insulin-dependent diabetics is greater than 100? |
| Step 1 | Obtain null and alternative hypothesis.  HO : The Average Percent Ideal Body Weight for insulin-dependent   diabetics is less than or equal to 100 HA : The Average Percent Ideal Body Weight for insulin-dependent   greater than 100 |
| Step 2 | Use a 5% an error rate which implies conclusions will be made with 95% confidence. |

|  |  |
| --- | --- |
| Step 3 | Completing the test in JMP…        Test Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_  Question: Is this value extreme? Discuss. |

|  |  |
| --- | --- |
| Step 4 | Determine p-value and make the statistical decision.    The Decision Rule: If the p-value is less than the error rate, then the data is said to support the alternative hypothesis.  P-Value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Statistical Decision: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Step 5 | Conclusion – Writing a final statement in laymen’s terms |
| Step 6 | Obtain the 95% confidence interval for this problem    What goes into getting this interval?    Sketch this interval on the line below.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Interpret the meaning of this 95% confidence interval.  JMP has the ability to do the “more appropriate” one-sided interval.   |  |  | | --- | --- | |  |  |   Sketch this interval on the number line below.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Interpret the meaning of this 95% confidence interval. |

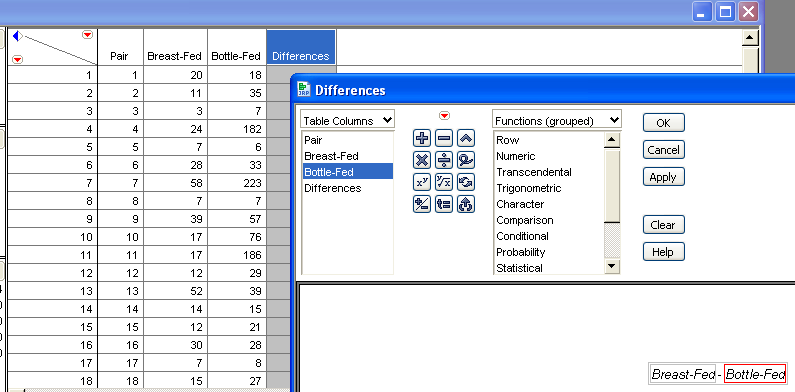
**Aside**: Investigating the effect of one outlier…



**Example 9.3** (Example 2.5, Rosner) A common symptom of otitis media (ear infection) in young children is the prolonged presence of fluid in the middle ear. The hypothesis has been proposed that babies who are breast-fed for at least 1 month may build up some immunity against the effects of the disease. A small study of 24 pairs of babies is set up, where the babies are matched on a one-to-one basis according to age, sex, socioeconomic status, and type of medications taken. One member of the matched pair is a breast-fed baby and the other was bottle-fed.

**Research Question:**  Is there a (statistical) significant difference in the duration of ear infection between the breast-fed and bottle-fed babies?

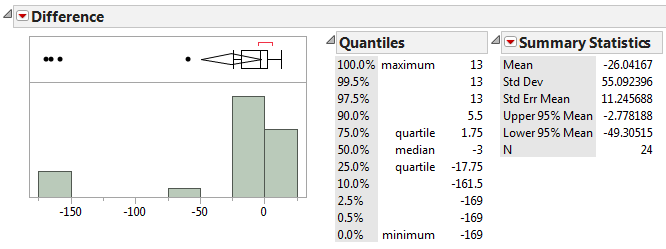
If comparisons are going to be made within each pair, then computing differences between the groups makes good sense. The differences can be computed in JMP automatically.



|  |  |
| --- | --- |
| Differences for each pair | Identifying the Number of Positives… |
|  |  |

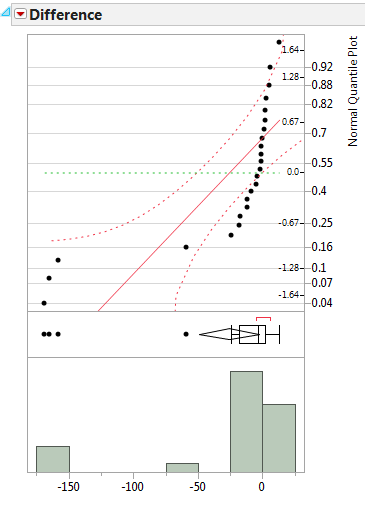
When we analyzed earlier in this course, we used the binomial distribution which allowed us to have only two outcomes. We do not have this restriction in Chapter 5, thus we can consider more than just the sign of the difference. In particular, we can consider the size of the difference which is useful because a difference of -2 is VERY different than a difference of -169!

Summary Statistics for the Differences…

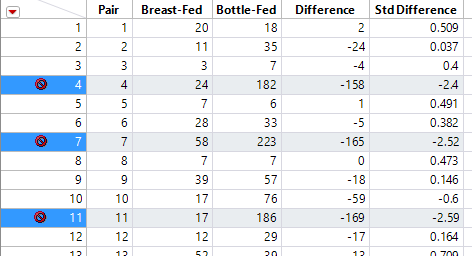


Does the observed data tend to support or refute the research question of differences existing? Discuss.

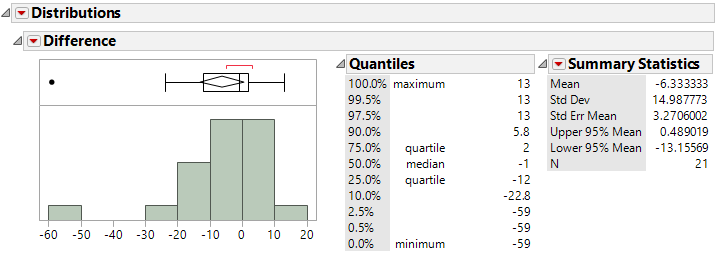
Checking the normality assumption (with all observations)



What happens if the outliers are removed from the data?



Descriptive statistics without the three outliers…

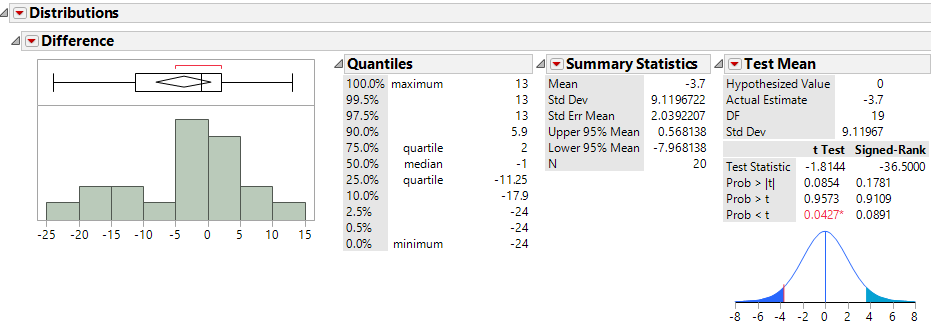


Discuss any differences here from what was observed with all 24 observations.

|  |  |
| --- | --- |
| Step 0 | **Research Question:**  Is there a (statistical) significant difference in the duration of ear infection between the breast-fed and bottle-fed babies? |
| Step 1 | Obtain null and alternative hypothesis.  HO : Average Duration for Breast-fed is the same as Bottle-fed HA : Average Duration for Breast-fed is different than Bottle-fed  or equivalently, |
| Step 2 | Use a 5% an error rate which implies conclusions will be made with 95% confidence. |
| Step 3 | Completing the test in JMP…   |  |  | | --- | --- | | All data (n=24) | Three outliers removed (n=21) |   Test Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_  Question: Is this value extreme? Discuss. |

|  |  |
| --- | --- |
| Step 4 | Determine p-value and make the statistical decision.  The Decision Rule: If the p-value is less than the error rate, then the data is said to support the alternative hypothesis.  P-Value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Statistical Decision: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Step 5 | Conclusion – Writing a final statement in laymen’s terms |
| Step 6 | Compute the appropriate 95% confidence interval for this problem   |  |  | | --- | --- | | All data (n=24) | Three outliers removed (n=21) |   Construct this interval on the line below.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Interpret the meaning of this 95% confidence interval. |

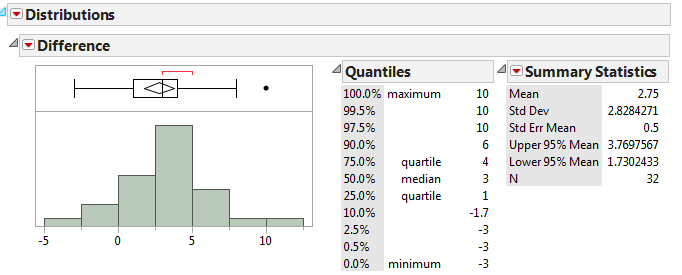
**Aside**: Removing the one additional outlier at -59 has a minimal impact…



**Example 9.4** Consider the degree of clinical agreement among two different physicians on their assessment of generalized lymphadenopathy. The data from this study is given here.

|  |  |
| --- | --- |
|  | Computing the differences for comparisons… |

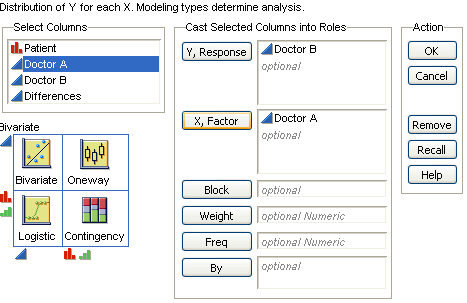
Summaries of the differences…



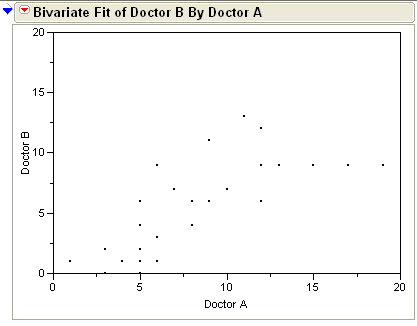
What can we say about the differences between these two doctors? Discuss…

|  |  |
| --- | --- |
| Step 0 | **Research Question:**  Is there a (statistical) significant difference exists between these two doctors? |
| Step 1 | Obtain null and alternative hypothesis.  HO : On average, no difference between doctors  HA : On average, difference exists between doctors  or equivalently, |
| Step 2 | Use a 5% an error rate which implies conclusions will be made with 95% confidence. |
| Step 3 | Completing the test in JMP…    Test Statistic: \_\_\_\_\_\_\_\_\_\_\_\_\_  Question: Is this value extreme? Discuss. |
| Step 4 | Determine p-value and make the statistical decision.  The Decision Rule: If the p-value is less than the error rate, then the data is said to support the alternative hypothesis.  P-Value: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Statistical Decision: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Step 5 | Conclusion – Writing a final statement in laymen’s terms |
| Step 6 | Compute the appropriate 95% confidence interval for this problem    Construct this interval on the line below.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Interpret the meaning of this 95% confidence interval. |

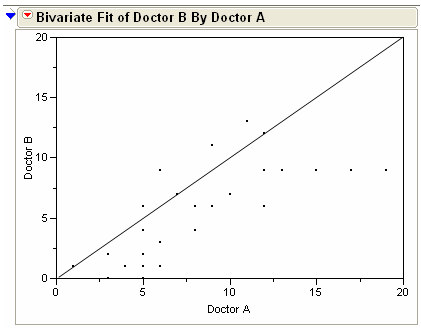
Next, consider creating a scatterplot (in JMP) to investigate the potential differences.



On this plot, we want to make sure the scales on the two axes match as a 1:1 relationship is expected.



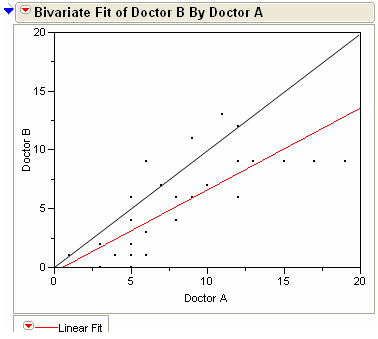
Adding a Y=X reference line to the plot which can be done under Fit Special…



**Questions**:

1. What does the Y=X reference line mean?
2. How many points fall exactly on the line? What does this mean?
3. Most of the points fall below the Y=X reference line. What does this tell you about the amount of agreement between these two doctors?

Adding the trend line (from the data) to this plot…



1. Notice that the trend line (from the data) and Y=X reference line start out about in the same spot. What does this tell us about the amount of agreement between these two doctors?
2. The trend line appears to be more flat than the Y=X reference line. What does this tell us about the amount of agreement between these two doctors?