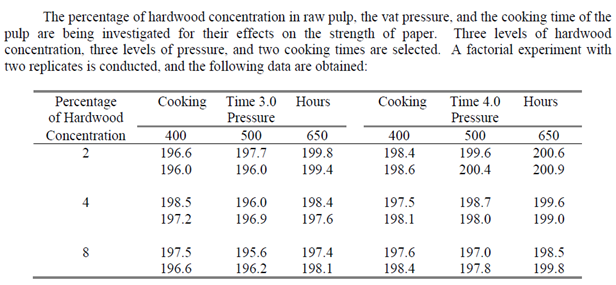
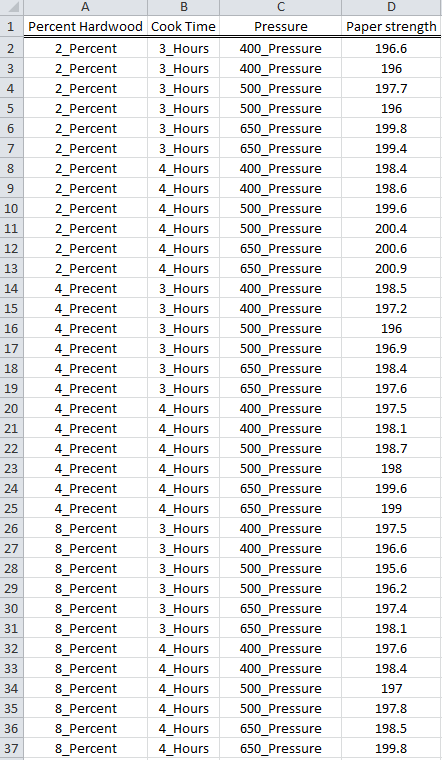
Handout #2: A Simple Comparative Experiment – Introduction to ANOVA

**Introduction**

We will begin with following ambitious example in which the relationship between many factors is considered.



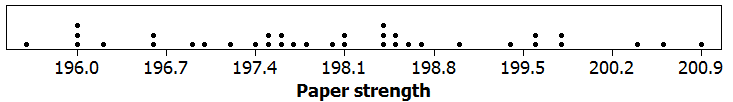
Putting this data into a useable (spreadsheet) format.



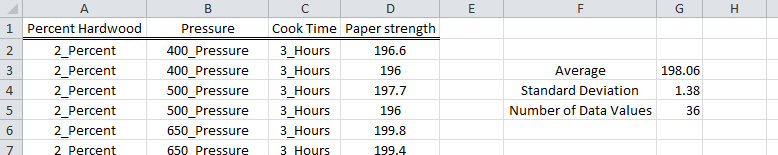
Understanding Paper Strength

1. What are the factors under consideration? What is the response under consideration?
2. What are the levels for each factor?

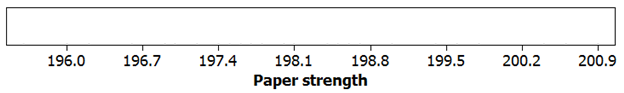
To begin, we will consider the response, paper strength. A dotplot of the paper strength without regard to any of the factors



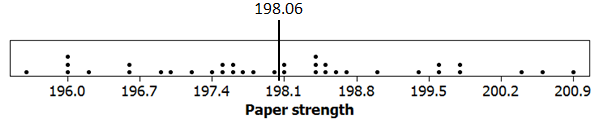
1. Use Excel to compute the average, standard deviation, and number of observations.



1. What would a set of data look like that had *less* standard deviation?



A visual depiction of standard deviation is shown here.



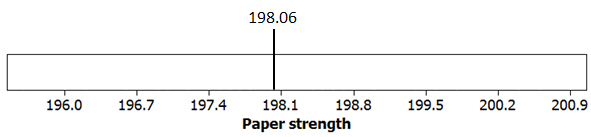
Often the standard deviation and the standard error are confused.

* Standard deviation: The degree to which *Individual* points vary from the mean

Formula for Standard Deviation:

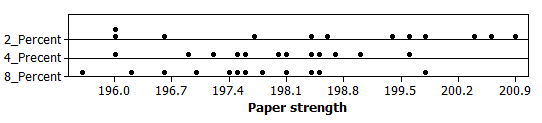
* Standard error: The degree to which an average will vary over repeated samples

Formula for Standard Error:

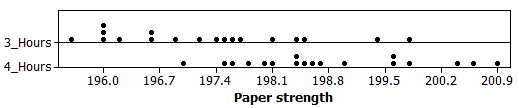


Understanding the Effect (Impact) of the Factors

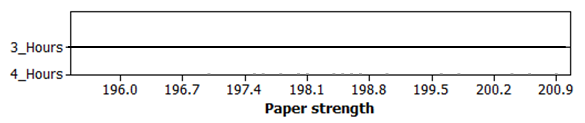
The impact of Percentage Hardwood



The impact of Cook Time



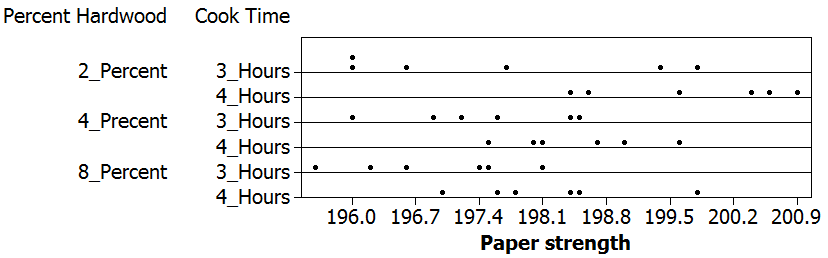
Sketch a plot in which Cook Time had even more impact than above



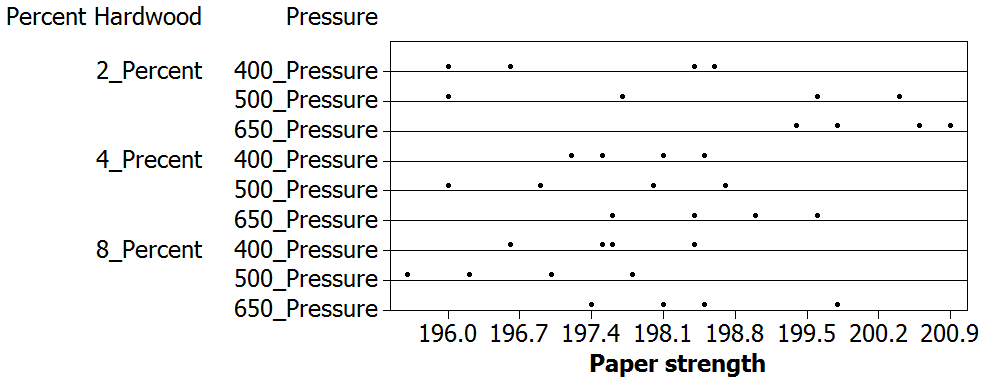
The impact of Pressure



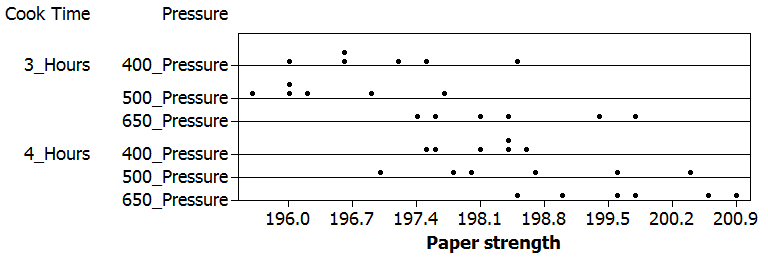
The impact of Percentage Hardwood and Cook Time



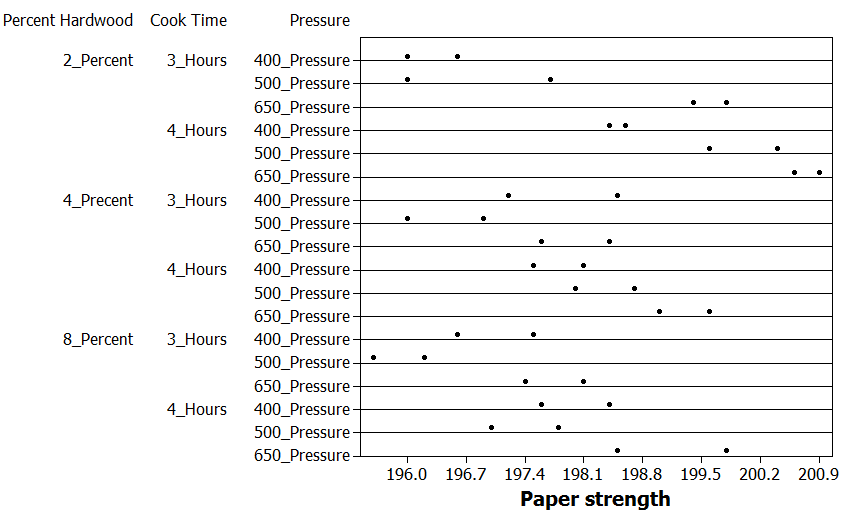
The impact of Percentage Hardwood and Pressure



The impact of Cook Time and Pressure



The impact of *all* factors



Before moving beyond our graphical analysis here, reconsider the impact of Percentage Hardwood and Pressure. The averages have been plotted on the dotchart.

|  |  |
| --- | --- |
| Simple dot chart with averages | Surface plot with averages |

You may find it useful to investigate certain slices of the surface.

Investigate the effect of pressure on paper strength when Slice: Hardwood = 2%.

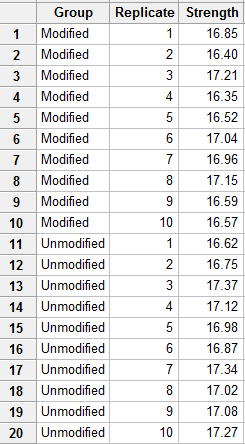
|  |  |
| --- | --- |
| Simple dot chart with averages | Surface plot with averages |

Investigate the effect of pressure on paper strength when Slice: Hardwood = 4%.

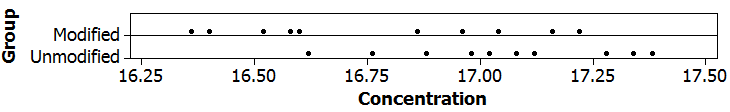
|  |  |
| --- | --- |
| Simple dot chart with averages | Surface plot with averages |

**The Statistical Approach**

Example 1.2: Consider the following data points from an experiment to determine the strength of portland cement mortar. These data points are from the Tension Bond Strength experiment discussed on page 24 of your text.



A dot plot of the data.

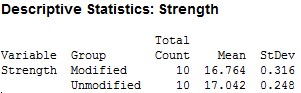


Questions:

1. Does the Modified formula seem to improve tension bond strength? Discuss.

1. What is the **factor** under consideration? What is the **response** under consideration?

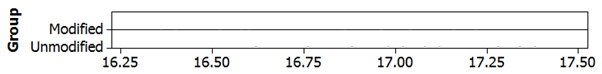
Summary statistics obtained from Minitab.



1. Compute the average for each group and sketch the averages on the number line below.

Average for Modified: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Average for Unmodified: \_\_\_\_\_\_\_\_\_\_\_\_



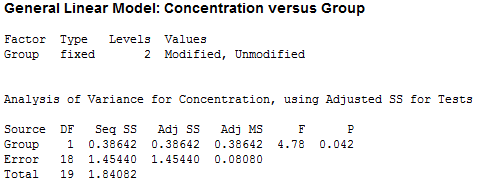
1. Consider the BIG QUESTION

BIG QUESTION: Can we say for certain that average for Unmodified will stay larger than the average for Modified over repeated outcomes of this experiment?

Answer: The ANOVA table.

**Minitab Version**

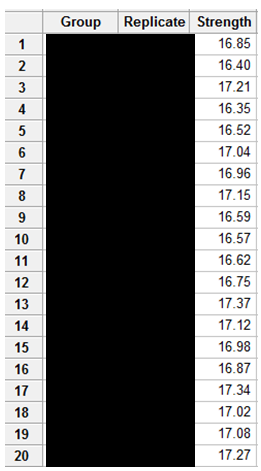
Stat > ANOVA > General Linear Model. Response=Concentration; Model = Group.



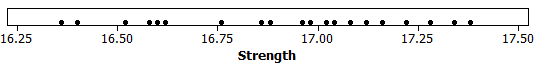
|  |
| --- |
| **JMP Version**  Analyze > Fit Model. Y=Concentration; Model Effects = Group |

**Measuring Error Assuming Factor is \*NOT\* Important**

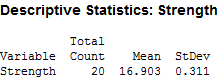
First, let’s consider the situation in which the factor “Group” had no effect. In this case, we would not expect there to be differences between the two groups; that is, we could ignore Group.



A sketch the data without regard to the factor level (i.e. group) is given here.



Similarly, the summary statistics without regard to the factor level.

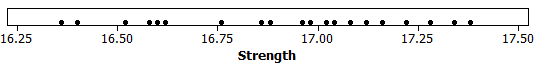


The average computed above is our best guess for the tension bond strength under the assumption that Group has no effect.

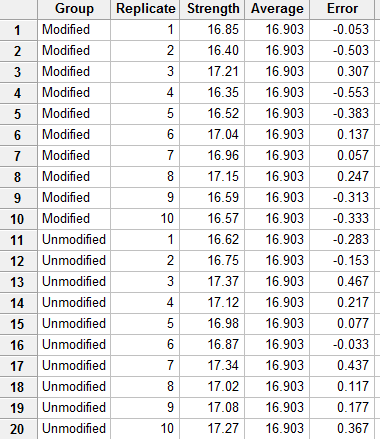
Average of all observations: \_\_\_\_\_\_\_\_\_\_\_\_\_

In statistics, we use the concept of **error** to determine which factors are significant. In particular, if the amount of error is reduced significantly when we consider the groups, then we say the groups are important and that the factor under consideration has some effect. On the other hand, if the amount of error is not reduced much by considering the factor, then the factor is said to be statistically unimportant.

Sketch the amount of “error” for each observation on the plot below, ignoring groups.



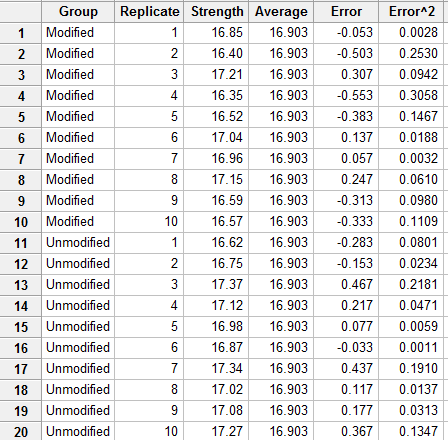
You can easily use Excel or Minitab to compute the error for each observation.



Questions:

1. What is the total amount of “error”?
2. What problems exist with computing a total amount of error in this manner?
3. How might we overcome this problem?

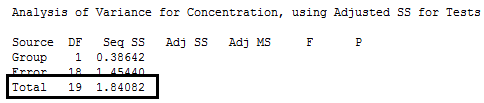
Getting rid of the canceling out effect by squaring





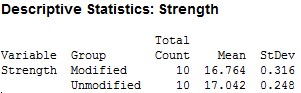
Total amount of error assuming group has NO effect: \_\_\_\_\_\_\_\_\_\_\_\_\_

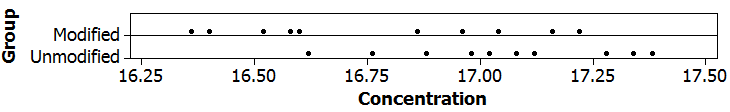
This quantity is given under SS Total in the ANOVA table.



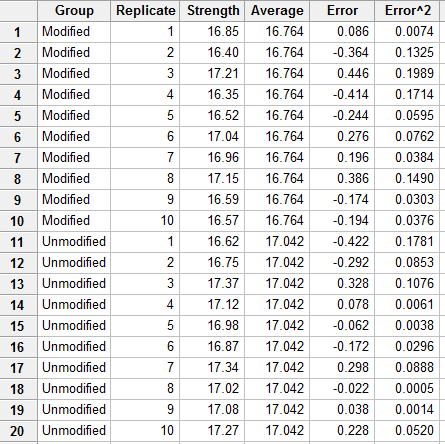
**Measuring Error Assuming Factor is Important**

Next, suppose the factor *does* indeed have an effect on strength. In this case, we would expect there to be differences between the two groups; that is, we should NOT ignore “Group”.

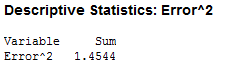




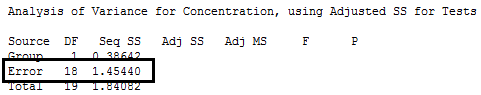
Again, we can compute the errors and square them in Excel or Minitab.



The total of the squared error values assuming the factor is important.



This quantity is given under SS Error in the ANOVA table.



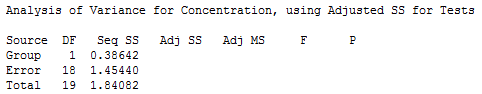
**Measuring The Importance of a Factor using Error**

What do we gain by considering groups? That is, how much is our total amount of error reduced after we account for a group effect?

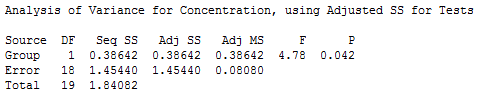
Difference: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Recall that if the amount of error is reduced significantly when we consider the groups, then we say the groups are important and that the factor under consideration has a significant effect. On the other hand, if the amount of error is not reduced much by considering the factor, then the factor is said to be statistically unimportant.

Examine the difference calculated above. Is this large enough to say that group has a significant statistical effect? To answer this question, we will use a statistical procedure called analysis of variance (ANOVA).



In the next section, we will consider the formulas for the remaining components of the ANOVA table.



**Measuring the Effect via the Formulas**

The name **analysis of variance** (ANOVA) is derived from a partitioning of the total variability into the following components. Suppose that yij represents the jth observation taken under factor level i. In general, we have *a* factor levels and *n* observations under the ith treatment.

Label the following observations using our notation:

Partitioning the Sums of Squares

Though we haven’t used the terminology or seen the formulas shown below, we have already calculated each of the following **sums of squares**.

|  |
| --- |
| 1. Total Sum of Squares  * SST = * This was calculated based on the assumption that groups had no effect.  1. Error Sum of Squares  * SSE = * This was calculated based on the assumption that groups had an effect.  1. Treatment Sum of Squares  * SSTrt = SST – SSE = * This represents the reduction in our measure of experimental error after accounting for group. |

Note that SST = SSTrt + SSE. That is, we have partitioned the Total Sum of Squares into two parts: the variation among factor level means (between treatments) and the variation due to experimental error (within treatments).

Degrees of Freedom

The degrees of freedom (df) associated with each sum of squares can be regarded as the number of independent elements in that sum of squares. The df for each sum of squares is shown below:

|  |
| --- |
| 1. For Total Sum of Squares: df =   We estimate one quantity from the data to find the total sum of squares (the overall mean); so, we lose 1 degree of freedom.   1. For Error Sum of Squares: df =   We estimate *a* quantities from the data to find the error sum of squares (a mean for each of the *a* factor levels); so we lose *a* degrees of freedom.   1. For Treatment Sum of Squares: df =   We start with *a* deviations and lose one degree of freedom for estimating the overall mean. |

Mean Squares

These are obtained by dividing each sum of squares by its associated degrees of freedom. Therefore, we have:

|  |
| --- |
| 1. Treatment Mean Square: MSTrt = 2. Mean Square Error: MSE = |

The F-statistic

|  |
| --- |
| F = |

Questions:

1. What does it mean if the F-statistic is large?
2. What does it mean if this ratio is small, say close to 1?

The p-value

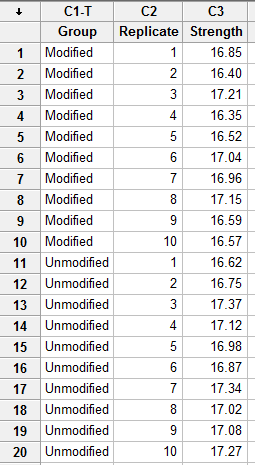
Recall that if the p-value is less than some predetermined error rate (we usually use an error rate of 5%), then the data is said to support the alternative hypothesis:

Ho: The factor has NO effect on the response variable.

Ha: The factor has an effect of the response variable.

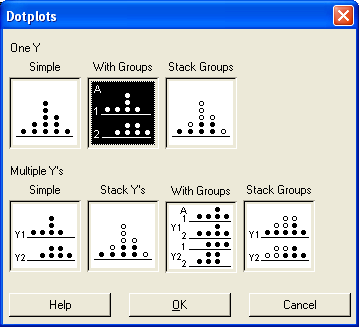
**The Analysis in Minitab**

First, important the data from Excel into Minitab.

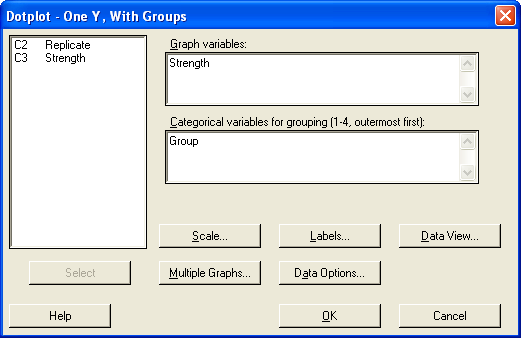


To construct dotplots, select **Graph > Dotplot…**

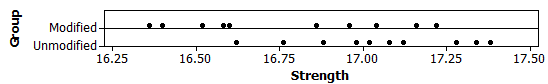
To make the dotplot with groups, choose the following:



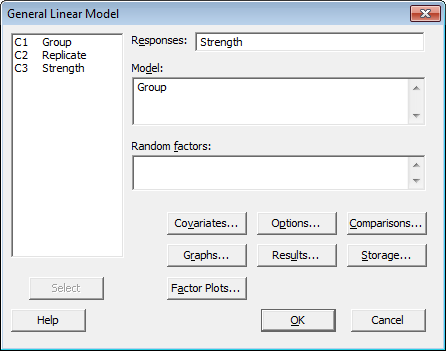
Enter the following in the next dialog box:



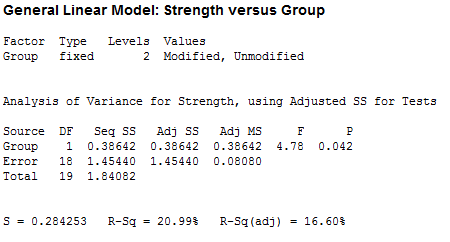
The following dot plot is given by Minitab.



To carry out the ANOVA in Minitab, select **Stat > ANOVA > General Linear Model**. Enter the following and then click OK.



The ANOVA output appears in the Session window:



Questions:

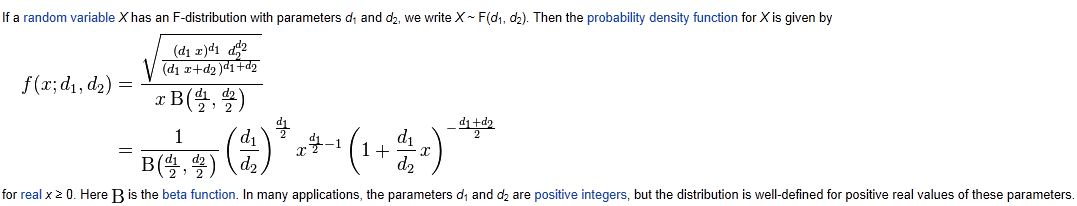
1. What is the total amount of error under the assumption the factor has no effect?
2. What is the total amount of error under the assumption the factor does indeed have an impact?
3. Explain how the following quantities were computed.

* DF Total =
* DF Error =
* DF Group =

1. What does the F-statistic represent? Explain.

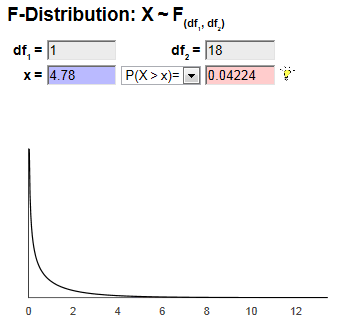
Obtaining the p-value.

* Wiki entry for F distribution



* There are free applets to compute this quantity. OR just let Minitab compute it!

<http://homepage.stat.uiowa.edu/~mbognar/applets/f.html>



The Decision Rule: If the p-value is less than 0.05, i.e. some predetermined error rate, then the data is said to support the alternative hypothesis.

Ho: The factor has NO effect on the response variable.

Ha: The factor has an effect of the response variable.

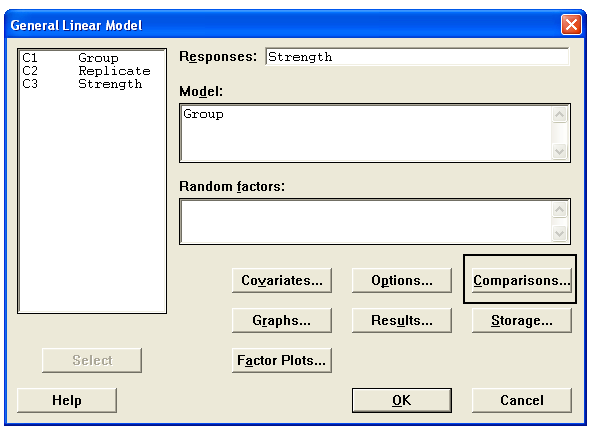
Question:

1. What can we (statistically) conclude regarding the importance of this factor?

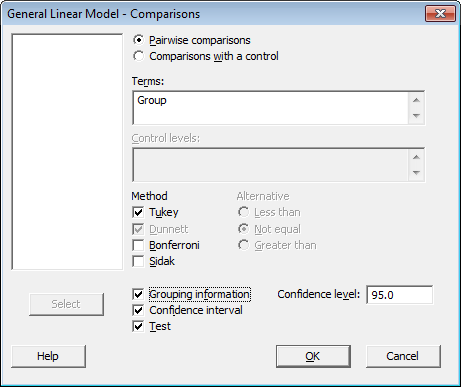
**Expressing the degree of the difference**

A **confidence interval** can be used to statistically measure the amount of the difference between two factor levels. That is, we can use a confidence interval to measure the true effect of a factor across its levels.

To compute a confidence interval in Minitab, select **Stat > ANOVA > General Linear Model**. The following window appears:

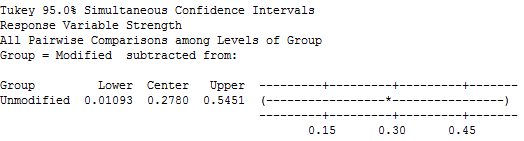


Select **Comparison**, and in the next window specify the following.



Click OK, and the output appears. This is divided into two sections, one for the confidence interval and one for the hypothesis test.

**Confidence Interval**



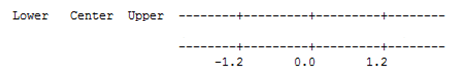
Questions:

1. Interpret this 95% confidence interval:
2. Why does it make sense that this confidence interval does not include zero? Discuss.

Note: Certainly, it is possible that a factor may not have an effect on the response. In this case, the p-value discussed above will not be small.

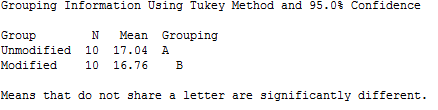
Questions:

1. On the number line below, provide a 95% confidence interval that would result when the p-value from our ANOVA analysis is not small.



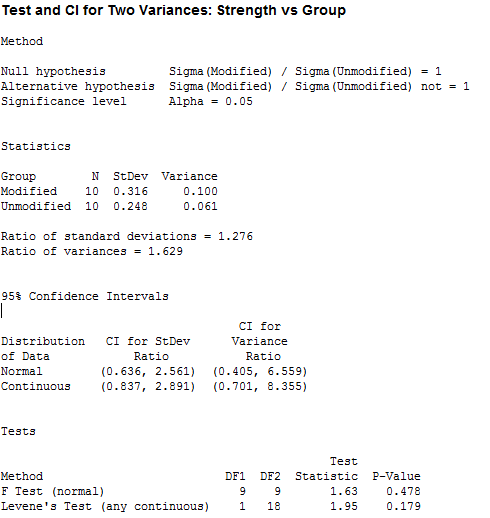
**Letterings**

One traditional approach for expressing which factor levels are different than others is to use letters. Groups which do not share letters are different. The factor under consideration here only has 2 factor levels; thus, we know these two are different from each other.

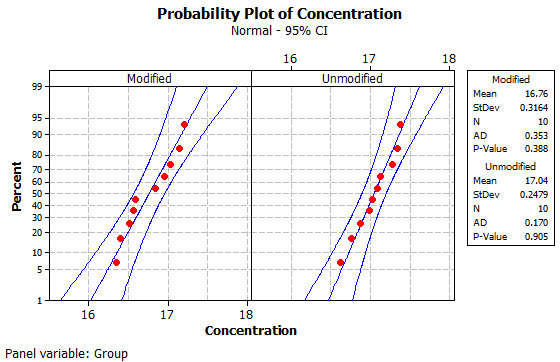


Assumption for Standard ANOVA Test

* Select Stat > ANOVA > Test for Equal Variances. Specify the Response and Factors to obtain the necessary output.



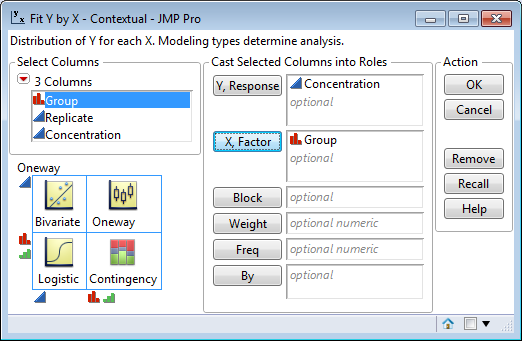
* Select Graph > Probability Plot. Place in the response in the Graph variables: box. Click Multiple Graphs… tab and specify Group under the By Variables tab.



**Appendix: The Analysis in JMP**

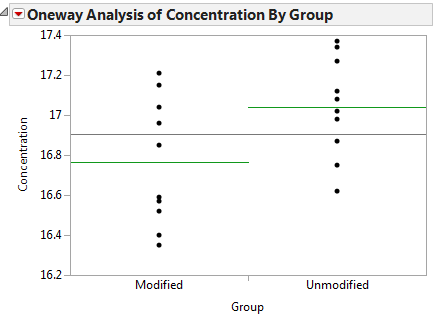
**Dotplots**

To construct dotplots, select **Analyze > Fit Y by X…**



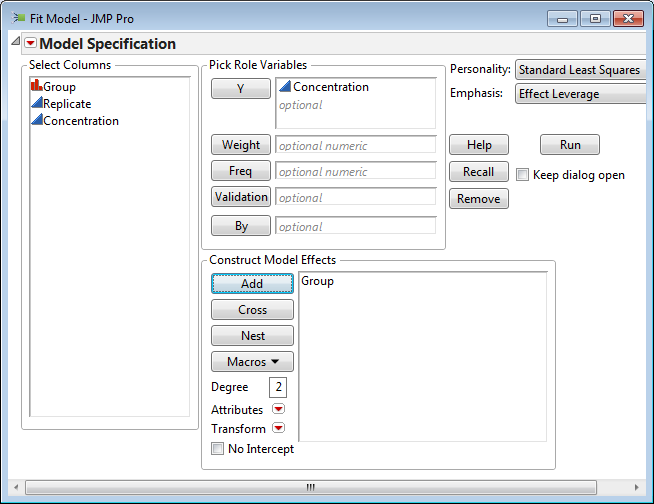
The default side-by-side dotplot is returned.

**Note**: There are several options available under Display Options from the Red Drop Down menu near the upper-left corner of the plot.

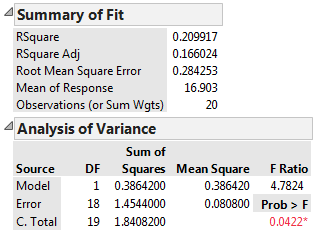


**ANOVA Output**

To carry out the ANOVA in JMP, select **Analyze > Fit Model** Enter the following and then click OK.

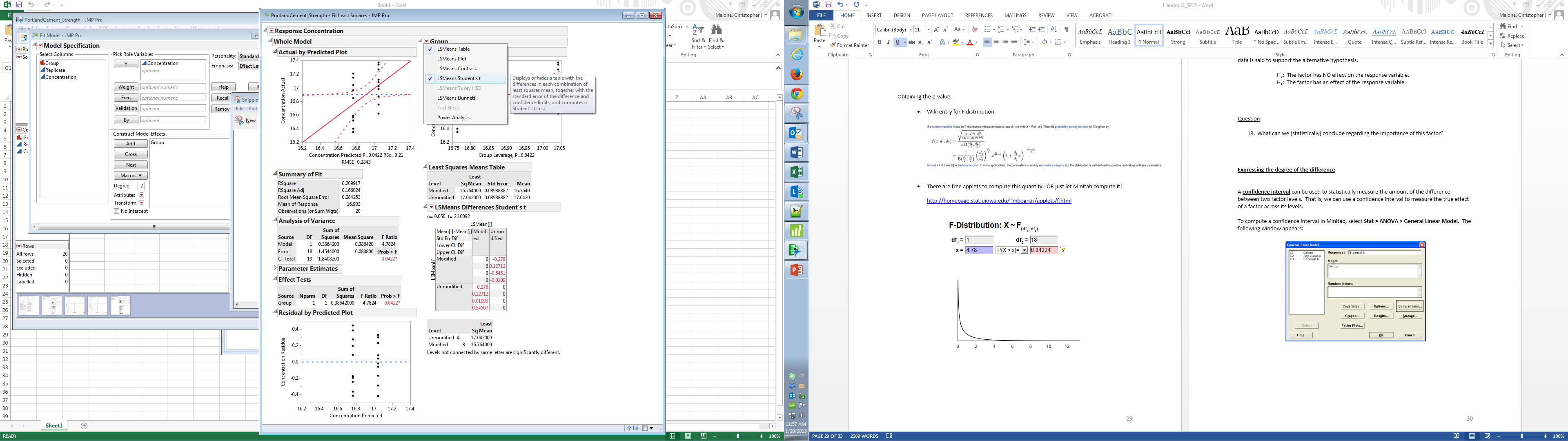


The ANOVA output appears in the Session window:

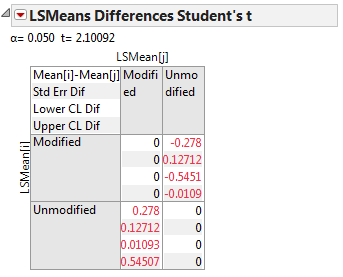


**Confidence Interval and Letterings**

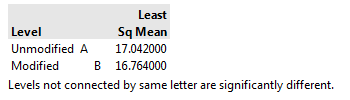
To compute a confidence interval in JMP, select **LSMeans Student’s t**  from the red drop down menu under Group.



Confidence Interval



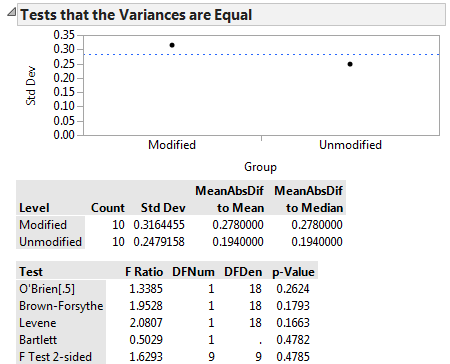
Letterings



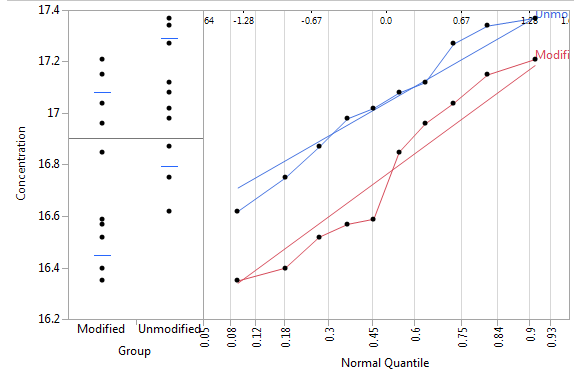
Assumption for Standard ANOVA Test

To obtain the necessary output for the assumptions in JMP, select Analyze > Fit Y by X. Place Concentration in the Y, Response box and Group in the X, Factor box.

* Select **Unequal Variances** from the red drop down menu for the Equal Variance tests.



* Select **Normal Quantile Plot** from the red drop down menu for the normality assumption.



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