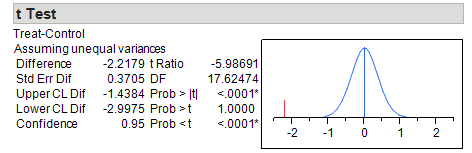
Kurt Medley  
Stats Final  
CPAT Spring 2013

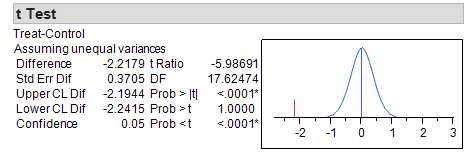
**Question 1 (10%).** T-test.

The x-axis represents the case number of the sample. The cases 1 through 13 represent the control group while the cases 14-25 represent the control treatment group. The outliers present in this grouping do not affect the usage of a t-test. There is not strong skewness in either direction.

Null Hypothesis: Malathion on oat plants has no effect on the beetle larvae population. Alternative Hypothesis: Malathion on oat plants reduces the population of beetle larvae and thus reduces the damage done by beetles.



PValue = < .0001. And we reject the null hypothesis.

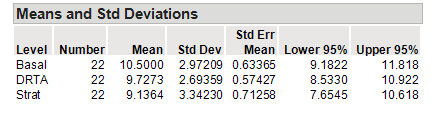


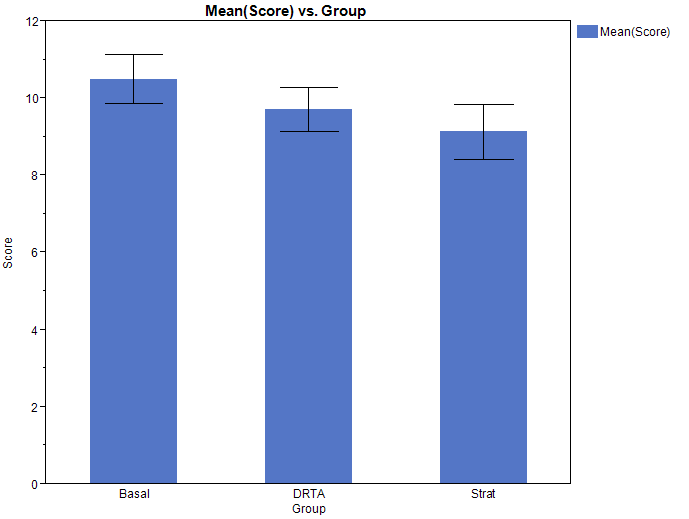
**Question 2 (10%).** Resampling to Compare 2 Means. Test your hypotheses for the above question using Resampling.

Running re-sampling stats, 0/1000 results gave an absolute value that was >= to the absolute value of the average beetle count of the control group minus the average beetle count of the treatment group. Pvalue = <.0001. It is obvious that the beetle count was not effected by mere chance. We reject the null hypothesis given these statistics.

**Question 3 (15%).** 1-way ANOVA.

Levene’s Test: PValue = .1947, this is > 0.5. There are no unequal variances.  
Shapiro Wilkes Goodness of Fit: 0.5893  
DRTA GOF : 0.0388  
STRAT GOF : 0.0727

We see that DRTA < 0.05; ANOVA re-sampling test in order: 



Null Hypothesis: There is no difference in the reading comprehension of students who have different methods of instruction.

Alternative Hypothesis: A method of instruction affects a student’s reading comprehension.

# of groups: 3 Groups

Sample Size for Each Group (N): 22

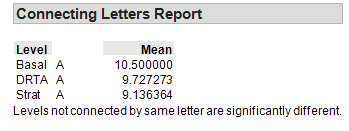
Mean (YBAR): 9.787879

Average of Each Group: Group 1: 10.5, Group 2: 9.727273, Group 3: 9.136364

Sum of Squares: 593.0303

SS Among: 20.57575

Shapiro Wilkes/GOF 0.0388. Resampling ANOVA is in order. SSAmong = 20.5. The repeat and scored 1000x. Counting the number of times the result was greater than the un-shuffled SSAmong stat divided by 1000 the PValue was roughly .3. This is a large PValue and so we fail to reject the null hypothesis. Reading comprehension scores are likely chance based and teaching methods had no roll.



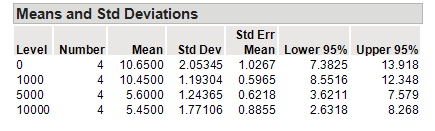
**Question 4 (15%).** 1-way ANOVA.

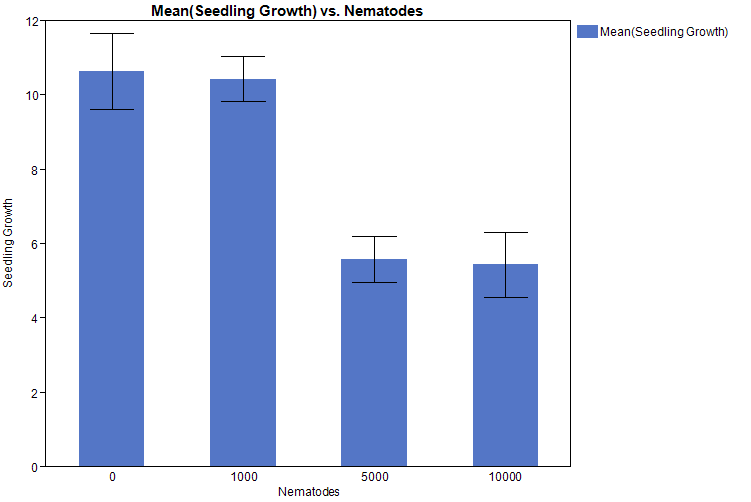
Levene’s Test for Unequal Variances:

Prob > F = 0.74

GOF: Pvalues for 0 nematodes: 0.24, 1000: .11, 5000: .24, 10000: .90

PValues > 0.05; And so probably from a normal population. Run ANOVA





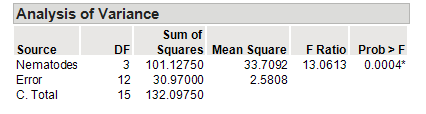
INCLUDES STD.ERR BARS

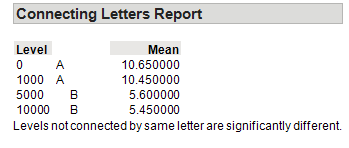
H0: The presence of Nematodes in the soil of plants have no effect on the Growth of seedling plants.

Ha: The presence of Nematodes in the soil of seedling plants reduces the growth of seedling plants.

ANOVA will compare the means of the each group and determine if it there is a significant difference between groups. It will give a p-value of how likely it is that the sample data occurred at random. If the p-value is very low, below 0.05 then we will reject the Null Hypothesis.

1. Run the ANOVA (in JMP or resampling) and report your results.





There is a statistical difference between groups 0, 1000 and 5000, 10000. There seems to be a relational difference between the number of nematodes in the soil and the growth of the seedling plants using the soil. Group 0 and 1000 have slight differences in means; group 5000 and 10000 also have slight variations in mean. Somewhere between 1000 and 5000 nematodes plant growth is altered.

**Question 5 (10%).** Linear Regression

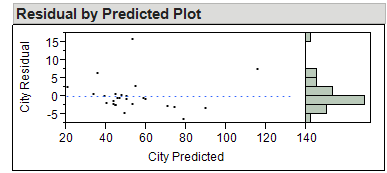
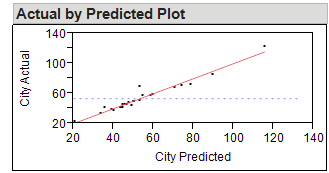
City = -2.580126 + 1.0935066\*Rural

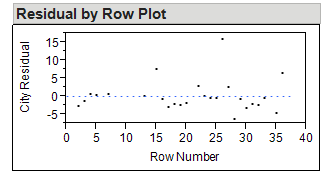
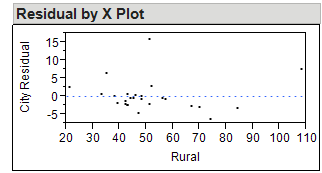
RSquare: 0.951128

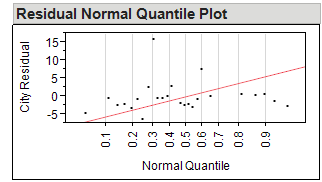
F Ratio: 467.07

Model: 1   
Error: 24

P-Value : <0.0001

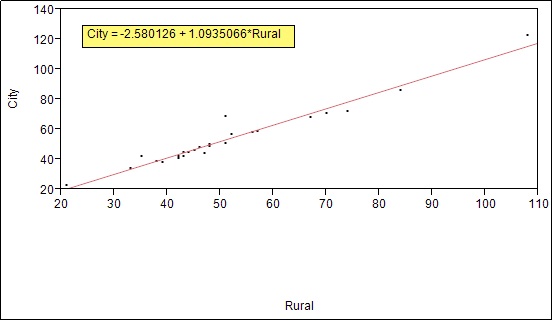
 



There is a Linear relationship between the dependent and independent variables. There does not seem to be a connection with errors.

Using the regression analysis we could predict the air pollution density of the city using the air pollution of Rural Area SW. The results of the analysis gives a Rsquared value of .95 and a Pvalue < .0001 which shows unlikely evidence of a linear relationship.

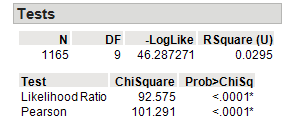
1. .

**Question 6 (10%).** Chi-Square.

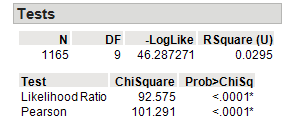
Nutrition and Illness share a relationship. Nutrition type may influence a particular illness or we may see an illness most associated with a nutritional category. These values may have association but they stand to be judged separately or one spawning the direct causality of another.

Null Hypothesis: There is no association of particular illness with a specific nutritional category nor can one assert that a nutritional type influences a specific illness.

Alternative Hypothesis: There is a strong association between a particular illness and a nutritional category; conversely, there can be nutritional category most associated with a particular illness.



DF: 9 (n – 1) \* (m – 1) n and m rows and columns.



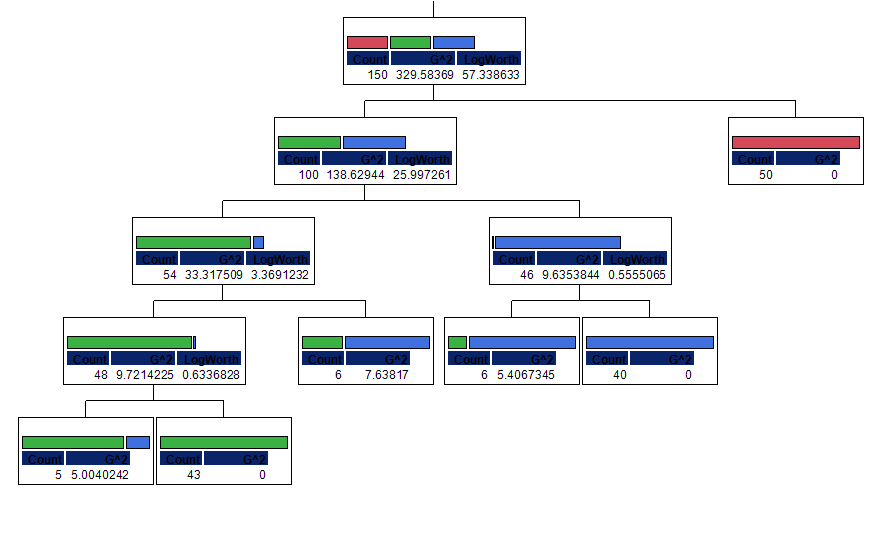
**(Optional) Question 7 (10%).** CART analysis. A data set from ML to be selected.

Iris Data set used in the ML lab

FIRST SPLIT Rsquared: 0.573

After the first split, the algorithm used the pedal length attribute.

After 5 splits, the Rsquared value was .930



**Question 8 (20%).**

The data I ascertained as being relevant from the tree spreadsheet were Site\_ID, Species\_ID, DBH, Live\_status and Height. I wanted to examine which species of trees had the greatest height. My analysis included the average of height of dead and alive trees. The dead trees had heights that affected the overall average of heights between the two categories. On average we saw that there was a greater percentage of living trees per site than dead trees. It might be logical to examine the evolutionary traits of PSME trees in the sites with heavy populations to theorize why those sites support that particular tree. Also it may be worth examining why site CH has the most dead ABAM trees.

After determining these questions a step in solving this mystery might involve particulars of the species in question. Does the density of other tree species on the site location limit the density of the PSME trees? Does the fact that site CH has the most dead ABAM trees associated with overcrowding?  
  
Null Hypothesis: The presence of other trees and potential overcrowding of other species of trees has no effect on the population of ABAM trees on site CH.  
Alternative Hypothesis: The presence of other trees on site CH has an affect on the population of ABAM tree density.   
 If there is an influx of alternative species of trees on site CH, it may affect the population of ABAM trees. We could examine growth patterns amongst these alternative species and determine if ABAM trees are ruled out by these patterns.

**Question 9 Short Essay**

I have limited experience using Microsoft Excel, so it was nice to use its functionality outside of summing up columns and rows of cells. The processes described in the lab activities made great usage of this tool. Using resampling stats was easy and intuitive using this program. Running other statistical tests was limiting (or my lack of knowledge made it limiting). Bar graph generation was not as intuitive as was in JMP. Exporting spreadsheets using .csv extension was great for analysis using R and Weka. The commonality of statistically pooled data seemed to have a universal extension in the .csv format. For straight observation of scientifically gathered data, Excel seems to be the tool to use.

I often found myself copying a spreadsheet from Excel into JMP for further analysis. JMP is a powerful statistical application that saves the manual application of function input that Excel forces a user to be familiar with. The libraries of relevant statistical tools JMP provides is significant. One could seamlessly transition from analyzing data using a one-way ANOVA to a Levenes and Fisher exact test. The graphs generated in JMP proved highly useful in data interpretation. Each instance of each test would be clearly marked and separated with its own frame. If one needed to run a series of common/practical statistical tests with accompanying graphics, JMP was definitely the tool to use.

I have experimented with R on several separate occasions but can vouch for its usefulness. Not only does it provide a great IDE for organization of user defined functions, it has a painless mechanism for importing large pools of categorical and numeric data using the .csv extension. Summary statistics are generated on the fly with the designation of which attributes needed. R requires a bit of patience for the non-programmer and programmer alike, but the functionality it offers is worth it.

**Question 10 Optional Short Essay.**

Techniques in machine learning require statistical analysis whereas I would say the converse is not necessarily true. Using the algorithms in Weka like OneR and J48 rely on a pool of statistical data. The data is used to generate rule sets for classifying further input. They generate a certain predicative quality based on these rules. I would say machine learning is an extension of the field of statistics, because it involves harvesting such data and analyzing it in a way machines can predict further input just like ANOVA is a technique for analyzing the variances of the means of collected data and drawing a conclusion.

That being said, I think there’s room in machine learning to produce new statistical tests. In that way, machine learning could help expand our reasoning about statistics as a field in itself. Being well versed in statistical analysis may help the machine learning practitioner rule out non-useful tests in the observation of and classification of data.