Bangor University
Laboratory Exercise 1, June 2008

Classroom Exercise
A forest land owner measures the outside bark diameters at 1.30 m above ground (called diameter at breast height or dbh) and total tree height from ground to tree tip for a sample of 20 trees on a small piece of land. The trees are equally spaces over the land area. The measures are:

Tree Number	Dbh (cm)	Height (m)
1	10.1	14.2
2	11.2	15.1
3	19.7	25.3
4	20.5	21.2
5	17.8	21.5
6	17.0	18.0
7	11.0	12.1
8	4.1	5.2
9	6.0	6.3
10	8.0	9.1
11	2.3	10.1
12	20.1	19.2
13	18.0	16.0
14	22.1	26.3
15	16.3	17.3
16	20.5	19.8
17	17.0	20.1
18	18.0	22.3
19	17.0	19.5
20	19.7	18.6

1. What is the name of the sampling design that was used? Explain why. Describe another sampling design that could have been used, and what the advantages/ disadvantages might be in using this alternative design.
2. Obtain a summary for the dbh variable, that includes:
    a. The sample mean
    b. The variance
    c. The standard error of the mean
    d. The mode
    e. The median
    f. The coefficient of variation as a percent
    g. A 95% confidence interval for the true mean (all of the trees).
    h. Given the sample data, and no assumptions about the probability distribution, what is the estimated probability that a tree will be more than 10.0 cm in dbh?

      i.   Given the sample data, and the assumption that it follows a normal distribution, what is the estimated probability that a tree will be more than 10.0 cm in dbh?

      j.   TRICKY question: If all trees were measured, how would the confidence interval for the mean be calculated?

3. Height is more difficult to measure than dbh (sometimes, in dense forest, particularly tropical forests, it is not possible to measure height without felling the tree or climbing it). Therefore, the land owner would like to obtain a height versus dbh equation. Then, dbh can be measured, and used to estimate tree height.

    a. Graph the height versus dbh for these sample data.

    b. Using this graph, decide if any transformations are needed, or if a simple linear equation using the original measured height and dbh is appropriate.

    c. If you do need transformations, make the transformation (try to transform dbh only, if possible), and graph height versus the transformed dbh (e.g., logarithm of dbh, reciprocal of dbh, or square of dbh, etc.). Decide if the transformation is appropriate (height versus the transformed variable are linear). If not, try other transformations until the relationship appears linear.

    d. Fit a simple linear regression of height versus your selected variable (dbh or a transformation of dbh) by hand. <u>NOTE: There is no need to change units to be the same for both variables.</u> Often variables are completely different units (e.g, biomass on one side and diameter on another side of the equation). Obtain (by hand):

        i.   The estimated intercept and slope. Use the estimated slope and intercept and overlay your equation over the selected graph in part c.

        ii.   Calculate the standard errors and 95% confidence intervals for the intercept and for the slope. Explain what these two confidence intervals mean.

        iii.   The coefficient of determination (r^2) and the standard error of the estimate (SE_E), also called the root mean squared error (Root MSE). What do these mean?

        iv.   Graph the fitted line over the original points.

        v.   Based on the graph, are the assumptions that the line fits the data and that variances of y's around the x's are equal met for your selected equation?

        vi.   How would you check whether the assumption that the errors are normally distributed around the regression line?

        vii.   How would you check the assumption that the observations are independent for these data?

        viii.   For a tree of 15 cm dbh, what is the predicted (estimated) height (i.e., mean height for a dbh of 15 cm)? Is this reliable? Calculate a 95% confidence interval for this estimated height.

4. Summarize your results for discussion in class.

Bangor University
Laboratory Exercise 2, June 2008

Computer Lab Exercise
Background:
R is a free software package that has been designed to analyze and graph data. A collection of people worldwide have developed underlined libraries of functions that you can use to analyze data. Because this is freeware, there may be "bugs" in the software. However, many of the library functions have now been tested by many users, and compared to other commercially available software packages such as SAS and SPSS.

To run R, first you need to load the software. This is fairly straightforward and has *already been done* on the computers in the lab. To load the software, you can go to R website directly. There are instructions for downloading software on the site. You will need to find a Cran near to your location, for faster service. A list of these can be found on the R website. This loads the standard package, with a number of libraries.

Useful things to know: When you run R, only some of the functions are brought into the work session automatically to save memory. To add others, you can use **require( )** where the package is given in brackets. Also, there are many other parts of R that are extra to the main package. To bring these in, you need to use **library( )** to access the website and add another library with a number of functions.

When you run R, you get a work session window, and you can also open script and get a separate window with the script. The script is just R commands, organized and put into a text file. Any lines that start with # are just *comments* that you can add to explain what the script does. Another window opens up if you get a graph using R.

R is case sensitive. This means that the variable Trees is not the same as the variable trees for example. Also, R does not like spaces nor special characters. Instead, use a '.'

Graphs: R is very good at graphs, using a function **plot( )**, where there are a number of arguments in the brackets (i.e., the x variable, the y variable, labels, type of graph, etc. However, only one graph appears at once, in a separate window. When you graph in R, you should save the graph (e.g., as a .jpeg file), before moving to the next graph. R can also do multiple graphs.

Help: The R website has a number of manuals that you might find useful. There are also a number of very useful books published by Springer, and Chapman and Hall publishers that I find very useful. On the outline for this course, you will find the website by Dr. Andrew Robinson, of Melbourne University, on an introduction to R.

At any time, you can also use **help( )** where the function is given in the brackets. This help is a bit hard to follow, and is really met to tell you the specific options for a function. However, there are also a few examples with the help that you might find useful as you are using R.

Exercise:
Using the same data for exercise 1, repeat all of the calculations and graphs using R. Script can be found in *exercise2_ht_dbh.R* file. You will run this in segments, by copying a part of the script into the work session, and then running that part OR by highlighting a part of the script and using **Ctrl**+**R** to run that part of the script. The work session window will include the R commands, and the outputs. At any time, you can copy and paste any part of the session window into a WORD file, or store the entire work session window.

As you do the parts of the exercise, compare the R results to your hand calculations. Also, try to read the R script, to see what is being done.

NOTE: You may have to modify the R code, depending upon what transformation you chose to use for dbh (i.e., to make the relationship linear). Try the code first, as it is written, before you alter it. ALSO – running it in segments will help you to understand what is going on, so you can alter R and use it for other research projects later on.

Computer Lab Exercise
Background:
Multiple linear regression uses more than one x-variable to predict the variable of interest, the y-variable. The x's can be several different variables that we have measured, or can be the originally measures variables, plus transformations of these variables. For example, we may use dbh and dbh squared to predict height, rather then just dbh or just dbh squared. In the case of the transformed variables, we are trying to meet the assumption that the linear model is correct.

Exercise:
Using the same data for exercise 1, run a multiple linear regression to predict height from dbh and transformations of dbh.   Script can be found in *exercise3_mlr.R* file
As you do the parts of the exercise, try to read the R script, to see what is being done.

For discussion:
For predicting height, do you need both dbh and another transformation of dbh (multiple linear regression)? OR is dbh sufficient (simple linear regression)? Or is just a transformation of dbh (simple linear model) sufficient?

Computer Lab Exercise
Background:
Stepwise methods can be helpful for selecting some x'variables for predicting the y-variable. Methods can be forward (in only), backward (out only) or both (in and out). The resulting subset of x variables can be different, depending upon the method used. Once subsets of x variables are obtained using these selection methods, a full regression can be run, and the assumptions checked, etc.

Exercise:
We will use the plot data found in *stand.txt* for this exercise. The data for each plot were compiled to obtain volume per ha, basal area per ha, stems per ha, top height, quadratic mean dbh, average age, site index. The script can be found in *exercise4_stepwise.R*.

Run the script in sections, as before, to be able to understand what the R code does. Then, using one of the subsets of selected variables, run a full regression analysis and check assumptions, etc.

For discussion:
How useful were these selection methods for choosing x variables to predict volume per ha? Did you obtain a good result with your full regression using the subset of x variables?

Computer Lab Exercise
Background:
R has some very useful graphics functions. These can be very helpful for conveying information to audiences in presentations and papers. We have already used histograms, and scatterplots for regression results.

Exercise:
We will use the tree data found in *trees.txt* for this exercise. There are 250 *Populus* trees and 250 *Abies* trees in this dataset. We will run some simple plots to visualize this fairly large dataset. The script can be found in *exercise5_plots.R*.

For discussion:
Which plot(s) did you find useful in visually describing these data?

Classroom & Computer Lab Exercise
A researcher wants to examine the impacts of thinning (tree removal) on growth of red pine trees in Ontario, Canada. There are three treatments: No removal (control), thinning (light – few trees are removed), heavy (many trees are removed). A plantation of 30 ha is selected, where trees are evenly spaced, with similar dbh's (diameter outside bark, measured at 1.3 m above ground) and are currently 15 years old. Fifteen areas are established in the plantation, each 1 ha in size (experimental unit). Each 1 ha area is then randomly assigned a treatment, resulting in five experimental units having each treatment. After 5 years, a number of 0.02 ha plots are established, systematically, over the each 1 ha area. The dbh's of all live trees are measured in each plot, and entered into an excel file. The average diameters are calculated for each 1 ha experimental unit resulting in the following values (data are in *exercise6_crd.txt*):

Treatment	Exp_unit	AveDbh
None	10	7.50
None	4	6.70
None	1	7.20
None	14	8.20
None	3	8.60
light	13	9.60
light	8	8.40
light	5	8.90
light	2	9.60
light	12	11.10
heavy	11	11.40
heavy	9	9.90
heavy	6	10.60
heavy	7	12.70
heavy	15	13.50

Using the script found in *exercise6_crd.R:*
1. Obtain a boxplot. Based on this boxplot, are there differences in AveDbh among the three treatments?
2. The null hypothesis is that there are no differences in mean of AveDbh among these three treatments?
   a. Check the assumptions by getting a histogram and normality plot of the residual values.
   b. If assumptions are met, set up your hypothesis (H0 and H1), obtain the F test statistic, the F critical value (or p-value), and make your decision (reject H0?), using the *lm* output. Use alpha=0.05.
3. Repeat the calculations in 2b BY HAND and compare to the R outputs.
4. Use pairs of means t-tests to check for differences between pairs of treatments. Remember to correct this test using a Bonferroni correction (i.e., divide alpha by the number of pairs of means).

Discussion:
Are these tests reliable? Were assumptions of Analysis of Variance met, or are transformations needed?

If assumptions were met, what are the results of your tests? Are there differences in AveDbh? If so, which thinning methods differ?

Computer Lab Exercise

In a second study, the impacts of thinning (tree removal) and fertilization on growth of red pine trees in Ontario are of interest. <u>The three levels for the first factor, thinning</u>, are: No removal (control), thinning (light – few trees are removed), heavy (many trees are removed). <u>For the second factor, fertilization, there are two levels</u>, from 1 (no fertilizer) to 2(fertilizer). In total, there are six treatments. Again, a plantation of 30 ha is selected, where trees are evenly spaced, with similar dbh's (diameter outside bark, measured at 1.3 m above ground) and are currently 15 years old. Twelve areas are established in the plantation, each 1 ha in size (experimental unit). Each 1 ha area is then randomly assigned a treatment, <u>resulting in two experimental units having each treatment</u>. After 5 years, a number of 0.02 ha plots are established, systematically, over the each 1 ha area. The dbh's of all live trees are measured in each plot, and entered into an excel file. The average diameters are calculated for each 1 ha experimental unit resulting in the following values (*exercise7_crd_two_factors.txt*):

Exp_unit	Thinning	FertLevel	ADbh
2	none	1	6.7
4	none	1	7.2
8	none	2	7.5
3	none	2	8.2
12	light	1	8.4
1	light	1	8.9
5	light	2	9.6
9	light	2	9.6
11	heavy	1	10.6
6	heavy	1	11.4
7	heavy	2	12.7
10	heavy	2	13.5

The researchers would like to know if there are differences in the mean of the AveDbh with different treatments.

1. Two analyses were run (*exercise7_crd_two_factors.R*. The first used AveDbh as the *y- variable,* and the second analysis used the logarithm of AveDbh instead. Which of these should be interpreted (NOTE: meets assumptions of equal variance and normality of residuals)?
2. Based on the analysis that met the assumptions, is there an interaction between the two factors? Use alpha=0.05.

THEN:

3. If there is an interaction, which treatments differ? (Use pairs of means t-tests at the treatment level, and remember to use the Bonfereonni correction.

OR:

3. If there is no interaction, does thinning change the diameter? In what way (higher or lower average diameter across thinning levels? (Use the pairs of means t-test for thinning levels – remember to use the Bonferonni correction).

4. If there is no interaction, does fertilizer change the average diameter? In what way? (Use the pairs of means t-test for fertilization levels – remember to use the Bonferonni correction).

Discussion:

Was the transformation of AveDbh needed?

Was there an interaction? If yes, which treatments differed?

If there was no interaction, is there a difference in mean of ADbh between thinning levels?

If there was no interaction, is there a difference in mean of ADbh between fertilization levels?

Computer Lab Exercise

A researcher sets up an experiment to the effects of different harvesting regimes on an understory plant called bunchberry (*Cornus canadensis*). This plant is quite small, with white flowers that form into a red berry. A forested area is selected and divided into three large areas that differ in the tree species composition, tree sizes, and tree density (Sites 1 to 3) since these conditions are expected to affect the amount of bunchberry. At each site, the researcher locates four areas of 0.5 ha each, surrounded by a buffer. Three treatments are randomly assigned to the 0.5 ha areas in each site: 1= no cutting; 2=partial cutting (1/3 of trees are removed); 3=partial cutting (2/3 of trees are removed); 4= all trees removed (Figure 1).

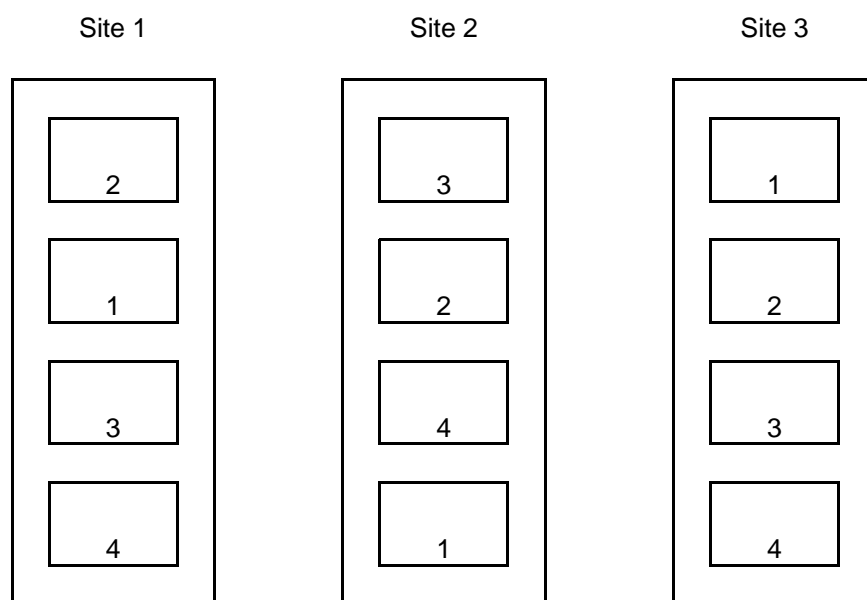| Site 1	Site 2	Site 3
2	3	1
1	2	2
3	4	3
4	1	4

Figure 1. Schematic of experiment to evaluate response of bunchberry to silvicultural treatments.

After four years, eight 2.07 m radius circular plots were laid out in a systematic pattern over each 0.5 ha area. In each circular plot, the number of bunchberry plants was recorded, along with reproduction success and growth variables for each plant. The researcher summarized the data to obtain the averages for each 0.5 ha area and passes this data on to you for analysis. You begin with an investigation for possible differences in the fruit weight between the treatments, based on the data in Table 1. (*exercise8_rcb.txt*)

Table 1. Fruit weight (g) by site and treatment.

Site	Treatment	Fruit_weight
1	none	0.013
1	one-third	0.022
1	two-thirds	0.034
1	all	0.051
2	none	0.032
2	one-third	0.043
2	two-thirds	0.053
2	all	0.075
3	none	0.053
3	one-third	0.065
3	two-thirds	0.073
3	all	0.095

Using R (*exercise8_rcb.R*), you create some boxplots, and calculate simple means for each site and for each treatment. You then use *lm* to look for differences in treatment means. Since the error terms may not be not be normally distributed, or there may be unequal variances, you try using the fruit weight in one analysis and using the logarithm of the fruit weight in another analysis.

Discussion:
1. Based on the box plots, what trends do you see over sites and over treatments?
2. What do you expect the i) number of observations; ii) df site; iii) df treatment; iv) df error; v) df total to be for this analysis? Verify that these match with the R outputs.
3. Is treatment a random or fixed effect? How about site?
4. For each analysis (y=fruit weight or y=log(fruit weight), assess whether the assumptions of i) independence of observations; ii) normality of residuals; and iii) equal variance across treatments is met using all available evidence.
5. Based on your assessment in Question 4, select one of these two analyses to interpret. For your selected analysis, test whether the treatment means differ, and, if they do differ, which means differ. **NOTE: Careful with the alpha levels for any pairs of means tests.**
6. For each treatment, calculate a 95% confidence interval for the true mean.
7. Can you test if the average of fruit weights differ across sites? Why or why not? **HINT: Is site a random or a fixed effect?**