Demonstration as a Case Study
Factorial ANOVA and Data compilation

All experiments begin with a Research Question, we are not focusing on that in this demonstration. The figures show the R code and output. The text with a hash and in green are the comments.

The Experiment

Dr Sally Campbell ran a trial experiment using two wheat cultivars (Cunningham and Sunvale) combined with and four salinity levels (0, 10, 20 and 30%). Each of the 4x2=8 treatment combinations has been replicated 3 times. We call one factor 'wheat' and the second factor is 'salinity'. We use names for the *levels* in these treatments.

Factor 1 Cunningham and Sunvale. These are variety names, and we have two levels.

Factor 2 0, 10, 20 and 30. These are percentage salt concentrations, we have four levels.

To run this experiment we determine the treatment combinations, and randomly apply these to the petri dishes, so that we have three replicates per treatment. A petri dish is an experimental unit. The allocation of treatments to Petri dish is assumed to be random within each replicate.

Measurements

The data presented here are the number of seeds (out of 20) germinated after 48 hours. Germination was considered to have taken place when the root was as long as the seed. **Note** we have clearly defined our measurement, both in the time and the acceptable germination criteria.

The results of our measurements are given in Figure 1 below.

We want to answer the following questions:

1. Do the cultivars generally germinate at the same rate?
2. Does increasing the salinity change the amount of germination?
3. (**Interaction**) Do the cultivars respond to increasing salinity in the same way?

CV	Salt conc%	Rep 1	Rep 2	Rep 3
Cunningham	0	17	17	18
Cunningham	10	8	14	12
Cunningham	20	2	8	5
Cunningham	30	3	1	3
Sunvale	0	19	16	16
Sunvale	10	15	15	11
Sunvale	20	4	11	10
Sunvale	30	4	5	3

Figure 1 Results of the number of wheat seeds germination per salinity

Figure 2 A petri dish at the conclusion of the experiment

Figure 3 The shoots were measured on a ruler at the end of the experiment

Rep	Cultivar	Salinity	Germination
1	Cungham	0	17
1	Cungham	10	8
1	Cungham	20	2
1	Cungham	30	3
1	Sunvale	0	19
1	Sunvale	10	15
1	Sunvale	20	4
1	Sunvale	30	4
2	Cungham	0	17
2	Cungham	10	14
2	Cungham	20	8
2	Cungham	30	1
2	Sunvale	0	16
2	Sunvale	10	15
2	Sunvale	20	11
2	Sunvale	30	5
3	Cungham	0	18
3	Cungham	10	12
3	Cungham	20	5
3	Cungham	30	3
3	Sunvale	0	16
3	Sunvale	10	11
3	Sunvale	20	10
3	Sunvale	30	3

Figure 4 The Germination data stacked in the correct way for entry into a spreadsheet

Data Exploration

Notice how the data needs to be entered into the statistical package. Each petri dish is identified in a row. We can type the data into Excel, and save as both as an Excel file and a csv file. If you have forgotten how to do this refer to the Excel notes in Module 0.

Use R to open a Project for your demonstration. This keeps all eth files in one Folder, and is easier to keep organised. Bringing the data into R studio from a file named **germ.csv** into R:

```
Germ <-   read.csv("germ.csv", header=TRUE)   # to read from the project directory
```

Check you have the correct number of variables etc.; using "str" and "name" in Rstudio.

Do basic data summaries (Figure 5) on the germination data. Before doing any analysis you need to find out as much as possible about your data; how does it behave (count/number, means, variances etc.). Load the Lattice library by typing in *library(lattice)* to enhance the graphics for RStudio then explore the data graphically by creating boxplots the germination (Figure 6 and Figure 7) and individual value plots for the Cultivars and Salinity (Figure 8, Figure 9 and Figure 10).

```
summary(Germ)  ### for Summary statistics

      Rep       Cultivar  Salinity  Germination
 Min.   :1   Cungham:12   0 :6     Min.   : 1.000
 1st Qu.:1   Sunvale:12   10:6     1st Qu.: 4.000
 Median :2                20:6     Median :10.500
 Mean   :2                30:6     Mean   : 9.875
 3rd Qu.:3                         3rd Qu.:15.250
 Max.   :3                         Max.   :19.000
 Germinat_perc      Cultivat
 Min.   : 5.00   Cungham:12
 1st Qu.:20.00   Sunvale:12
 Median :52.50
 Mean   :49.38
 3rd Qu.:76.25
 Max.   :95.00
```

Figure 5 The Summary Statistics for Germination data

```
###boxplots for cultivar per salinity
boxplot(Germination~Cultivar*Salinity,data=Germ,main="Germination",  ylab="Seed count",
xlab = "Cultivar/Salinity")
```

Figure 6 Syntax for Boxplot for Germination data (use # to make notes and describe the code/ syntax; green printing)

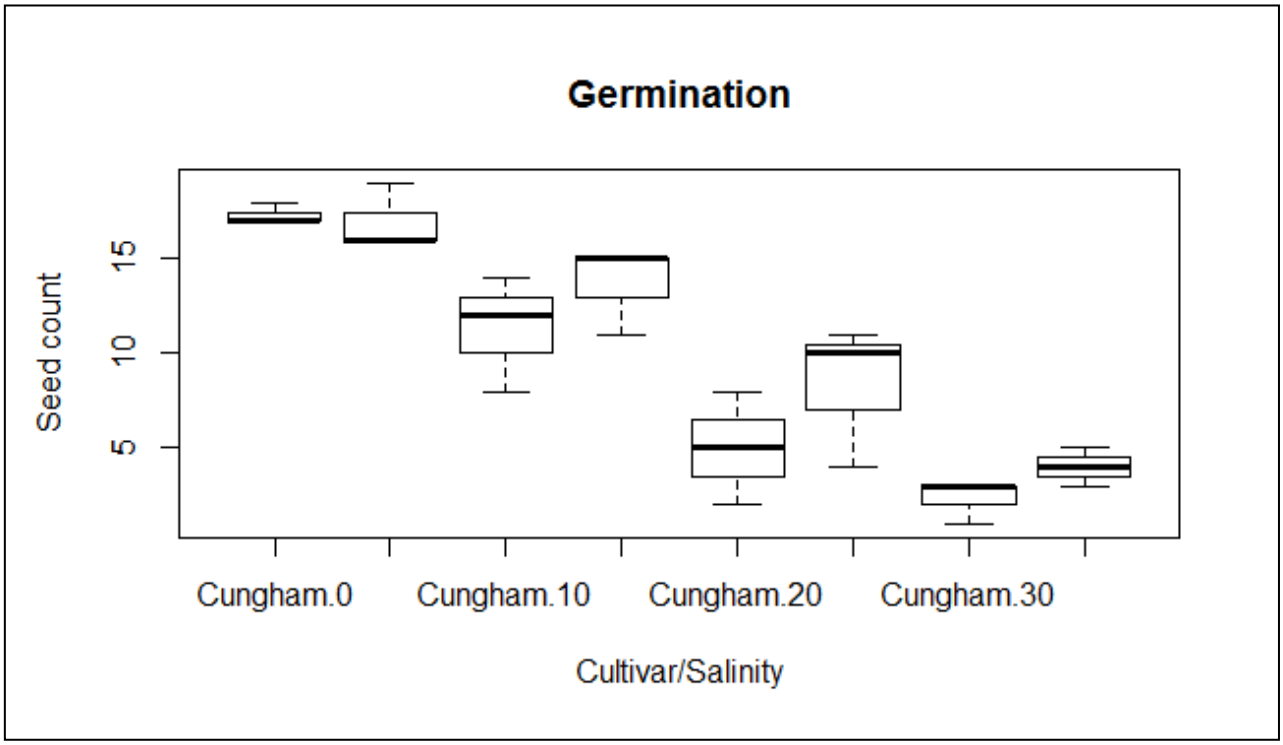

Figure 7 Boxplot of Germination data

```
#####plots individual for Cultivars

plot(Germ$Germination,Germ$Cultivar, ylab = "Cultivar", xlab ="Germination", pch=16)
#####plots individual for Salinity

plot(Germ$Germination,Germ$Salinity, ylab = "Salinity", xlab ="Germination", pch=16)
```

Figure 8 Syntax for individual value plots

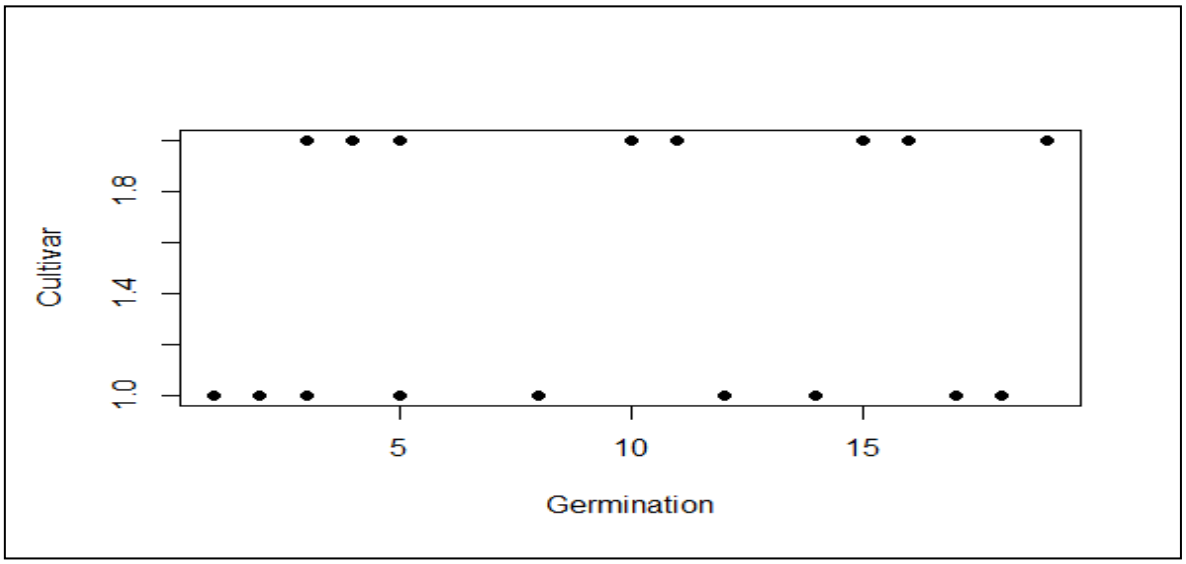

Figure 9 Individual Value plot of Germination per the Cultivar

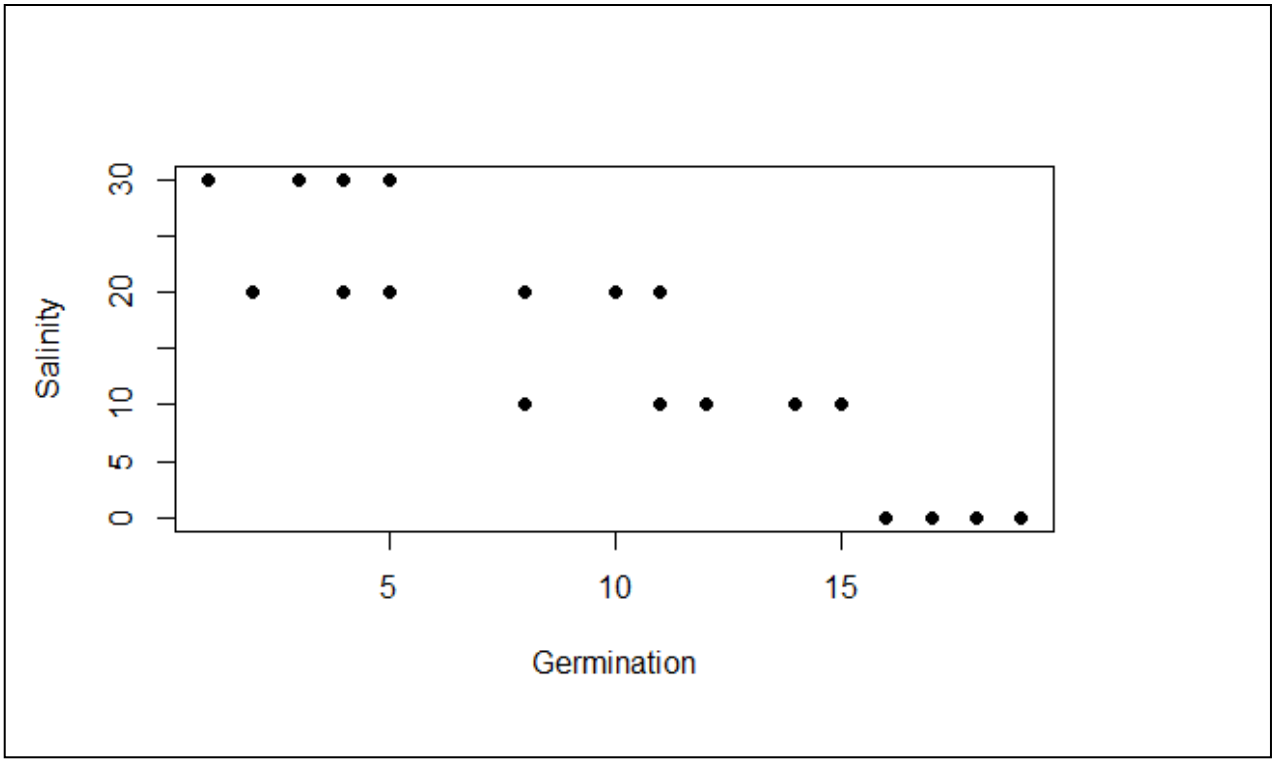

Figure 10 Individual Value plot of Germination per Salinity levels

All the plots above show the two cultivars responding markedly to increased salinity, with Cunningham showing a greater decrease in germination at the high salinity levels. As expected, as salinity levels increase there is a decrease in the germination rate of both cultivars.

ANOVA Data Analysis

Finally, we are ready to carry out an analysis of variance (ANOVA). The *Terms* to be tested are 'Cultivar' and 'Salinity' and the model needs to be entered as 'Cultivar + Salinity + Cultivar*Salinity' or 'Cultivar*Salinity'. The *Response* is 'Germination' or 'Germin_perc' and the data is 'Germ'. NB the blue printing is the command for the analysis.

```
###ANOVA for Germ without interactions, germ.aov is the object name for the ANOVA
germ.aov <- aov(Germination~Cultivar+Salinity, data =Germ)
summary(germ.aov)
## this is needed to 'print' the ANOVA table
```

Figure 11 Syntax for ANOVA without an interaction term

```
###ANOVA for Germ without interactions, germ.aov is the object name for the ANOVA
germ.aov<-aov(Germination~Cultivar+Salinity, data =Germ)
summary(germ.aov)  ## this is needed to 'print' the ANOVA table
            Df Sum Sq Mean Sq F value   Pr(>F)
Cultivar     1   18.4   18.38   3.558   0.0746 .
Salinity     3  692.1  230.71  44.672 8.42e-09 ***
Residuals   19   98.1    5.16
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 12 The ANOVA output without an interaction term

```
###ANOVA for Germ, germ.aov = object name for the ANOVA
germNI.aov <- aov(Germination~Cultivar+Salinity+Cultivar*Salinity,
data =Germ)
summary(germNI.aov)
## this is needed to 'print' the ANOVA table
                 Df Sum Sq Mean Sq F value   Pr(>F)
Cultivar          1   18.4   18.38   3.366   0.0852 .
Salinity          3  692.1  230.71  42.267 7.87e-08 ***
Cultivar:Salinity 3   10.8    3.60   0.659   0.5891
Residuals        16   87.3    5.46
---
```

Figure 13 The ANOVA output for Germ with an interaction term

```
#### to plot interaction first run this part be careful to get the syntax correct!!!

require(tables) ## need to load this each time you open RStudio
tabular((interaction (Cultivar, Salinity)+1)~(n=1)
      + Format (digits=3)*(Germination)*(mean + sd), data=Germ)
```

```
                                  Germination
interaction(Cultivar, Salinity) n  mean        sd
Cungham.0                        3 17.333       0.577
Sunvale.0                        3 17.000       1.732
Cungham.10                       3 11.333       3.055
Sunvale.10                       3 13.667       2.309
Cungham.20                       3  5.000       3.000
Sunvale.20                       3  8.333       3.786
Cungham.30                       3  2.333       1.155
Sunvale.30                       3  4.000       1.000
All                             24  9.875       5.929
```

```
##then plot the interaction plot using "with"
with(Germ, interaction.plot(Cultivar, Salinity, Germination))
```

Figure 14Syntax for Interaction plots using the pervious syntax model

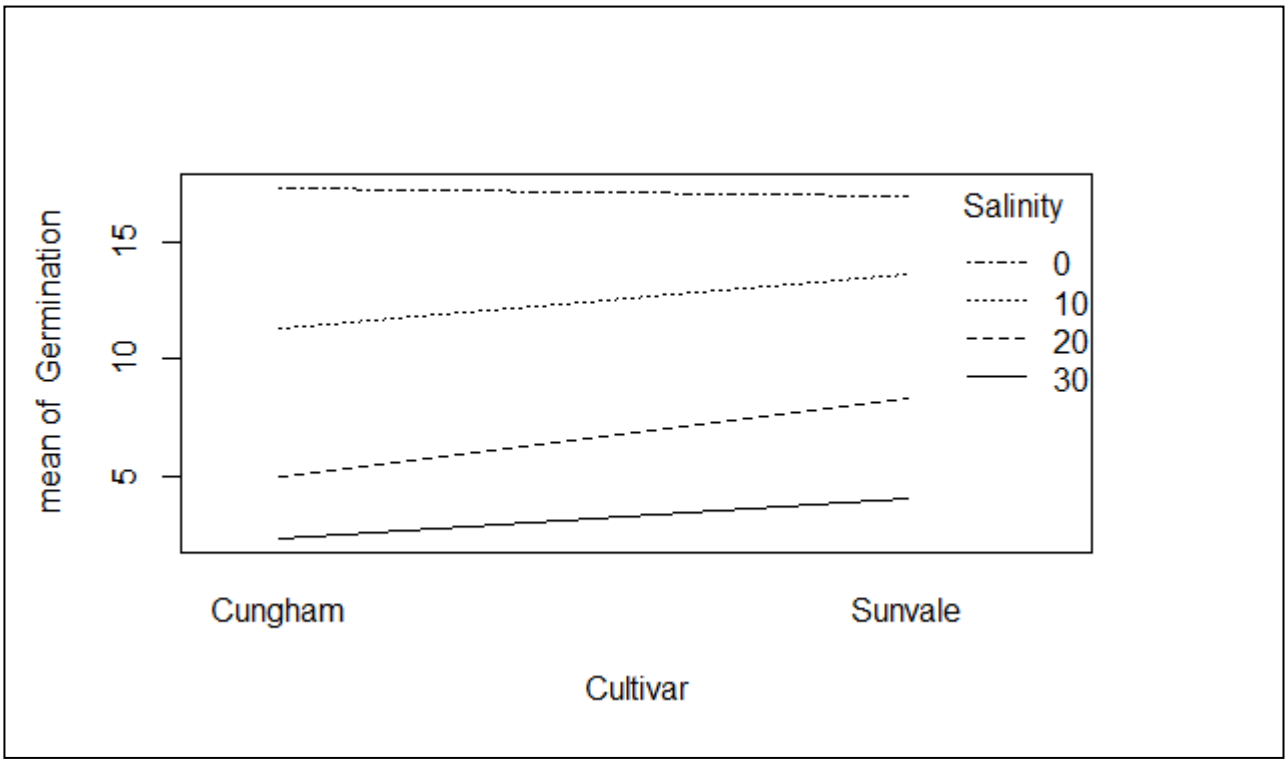

Figure 15 Interaction Plot for Cultivar for Salinity

```
###An alternative way this need you to load "require(effects)"
require(effects)
plot(allEffects(germNI.aov))
```

Figure 16 Another way of creating interaction plots

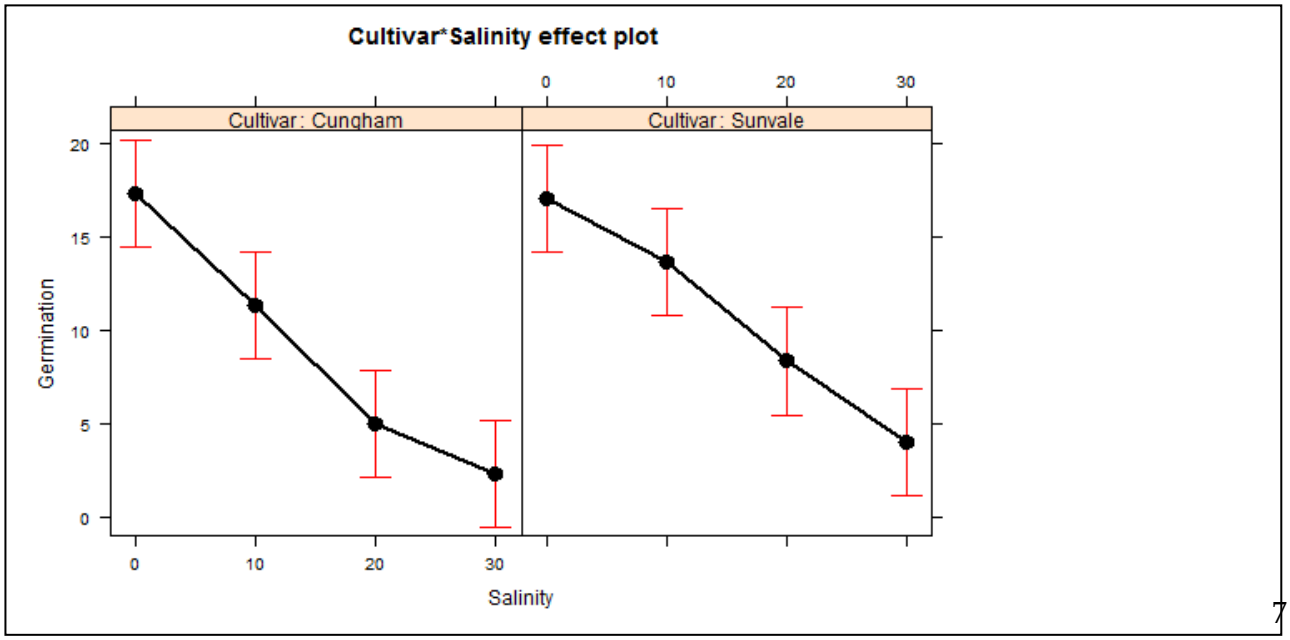

Figure 17 Interaction plots for Cultivar by Salinity levels

```
##another way to plot an interaction
xyplot(Germination~Salinity, group= Cultivar, data = Germ,
      auto.key = TRUE, type= c("p","a"))
```

Figure 18 Syntax for an interaction plot which only requires the lattice library to be loaded

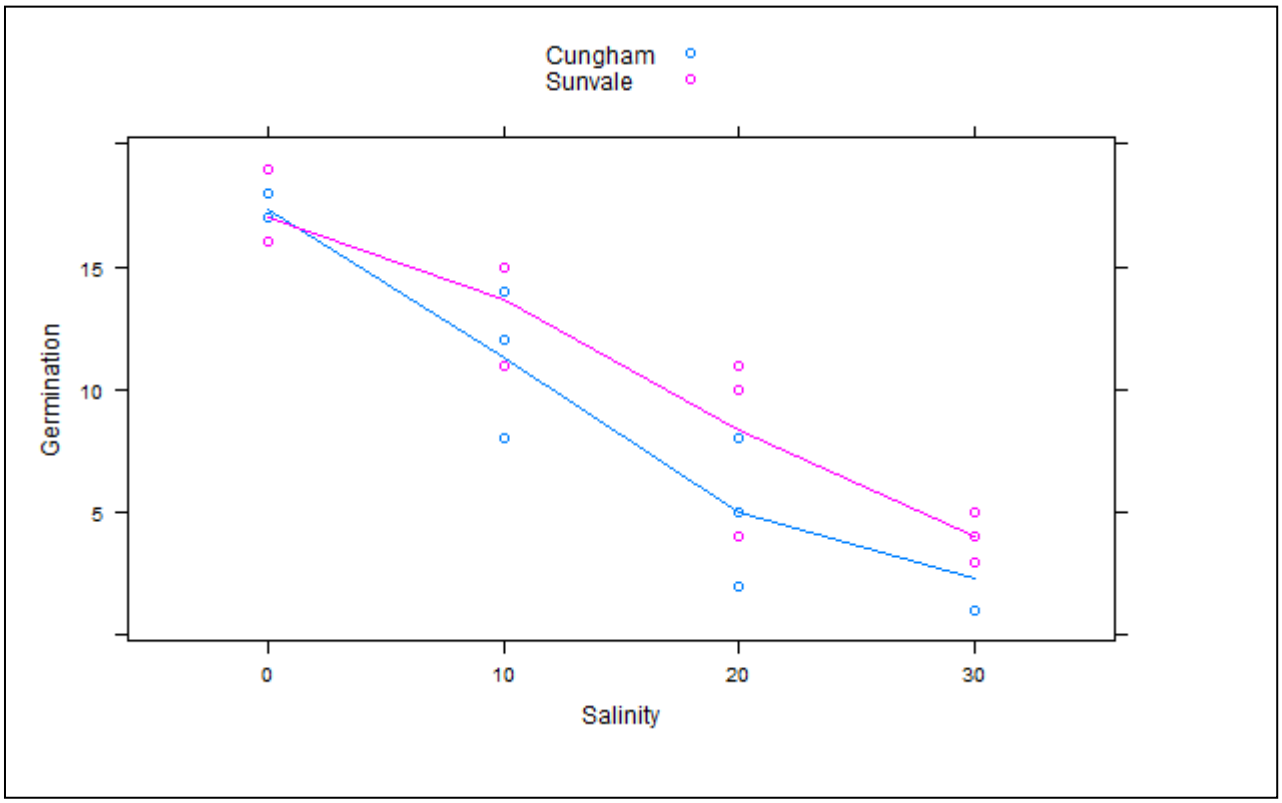

Figure 19 The interaction plot for the salinity (x axis) by varieties (as the two coloured symbols)

```
### model.tables gives grand mean plus the means for each cultivar
## and salinity level plus the means of each cultivar for each level
##of Salinity

model.tables(germ.aov,"means")
```

Figure 20 Syntax for means from the factors in the ANOVA

```
Tables of means
Grand mean

9.875

 Cultivar
Cultivar
Cungham Sunvale
   9.00   10.75

 Salinity
Salinity
     0     10     20     30
17.167 12.500  6.667  3.167

 Cultivar:Salinity
         Salinity
Cultivar  0      10     20     30
  Cungham 17.333 11.333  5.000  2.333
  Sunvale 17.000 13.667  8.333  4.000
```

Figure 21 The table of means for the factors in the ANOVA

A useful graphic when we have a factorial treatment structure is an Interaction Plot Germination is the *Response* and the *Factors* are Cultivar and Salinity (Figure 28)
Interaction plot similar to the scatterplot, but only showing the means, with connect lines between them. Looked at this way (means only), Cunningham shows slightly lower germination on average, except at '0' level (control level). You could suspect there might be an interaction at levels 10 and 20 but you need to analysis the data and use a multiple comparison test *only* if the interaction term is significant.

```
### this give an interaction plot of how the cultivars behave for
## each salinity level(dot, "p" and line, "a")

xyplot(Germination ~ Cultivar, group=Salinity, data=Germ,
       auto.key=TRUE, type =c("p","a"))
```

Figure 22 The syntax for the interaction plot for the Salinity levels

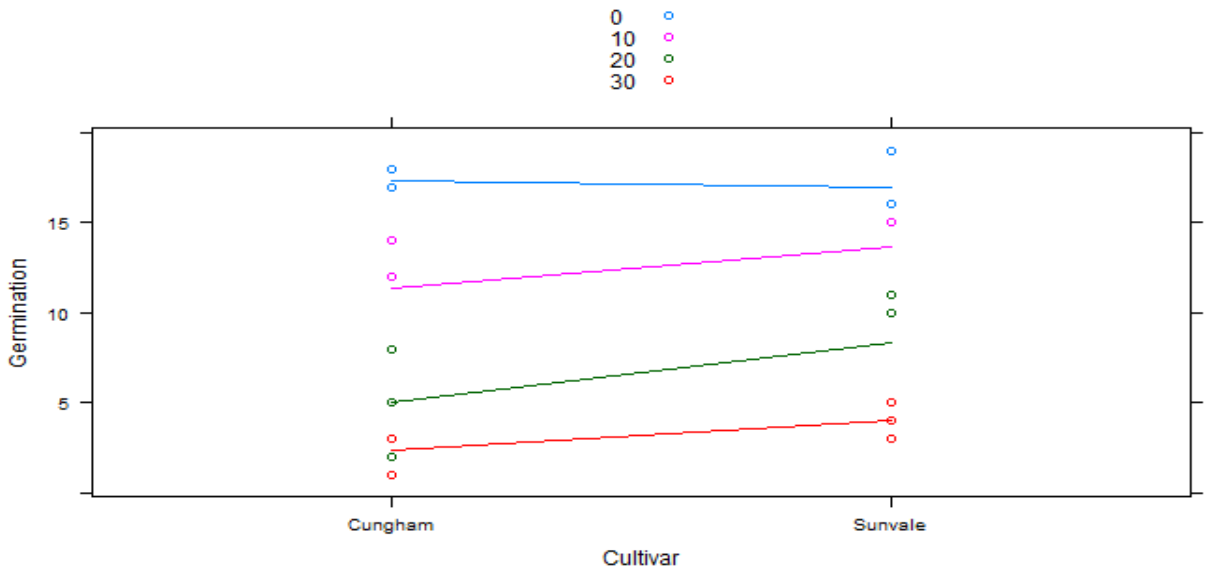

Figure 23 The interaction plot Cultivars for the Salinity levels

The three Assumptions for ANOVA

Remember the three assumptions for an ANOVA to be validly used? One is that the variances are equal another is that the residuals are normally distributed. Residuals can be generated and stored as a column (from the ANOVA already performed) in RStudio so that they can be plotted to check the assumptions (Figure 16). If the first assumption (equal variance) is reasonable, the variability should be consistent across all Cultivars and Salinity. The points need to be scatter randomly across the plots with no obvious pattern showing. Plot the 'xyplot' to check this. For the second assumption that the residuals are from a normal distribution have to be generated

```
###Add a column to the data Frame to store the residuals from the ANOVA.

Germ$residuals = residuals(aov(Germination~Cultivar+Salinity+Cultivar*Salinity,
data =Germ))
```

Figure 24 Generating a column of the residuals from the ANOVA

```
###Use a 'xyplot' to check for equal variances

xyplot(residuals ~ Cultivar | Salinity, layout=c(3,1), data=Germ)
```

Figure 25 The syntax for the scatter plot of the Cultivars per Salinity level

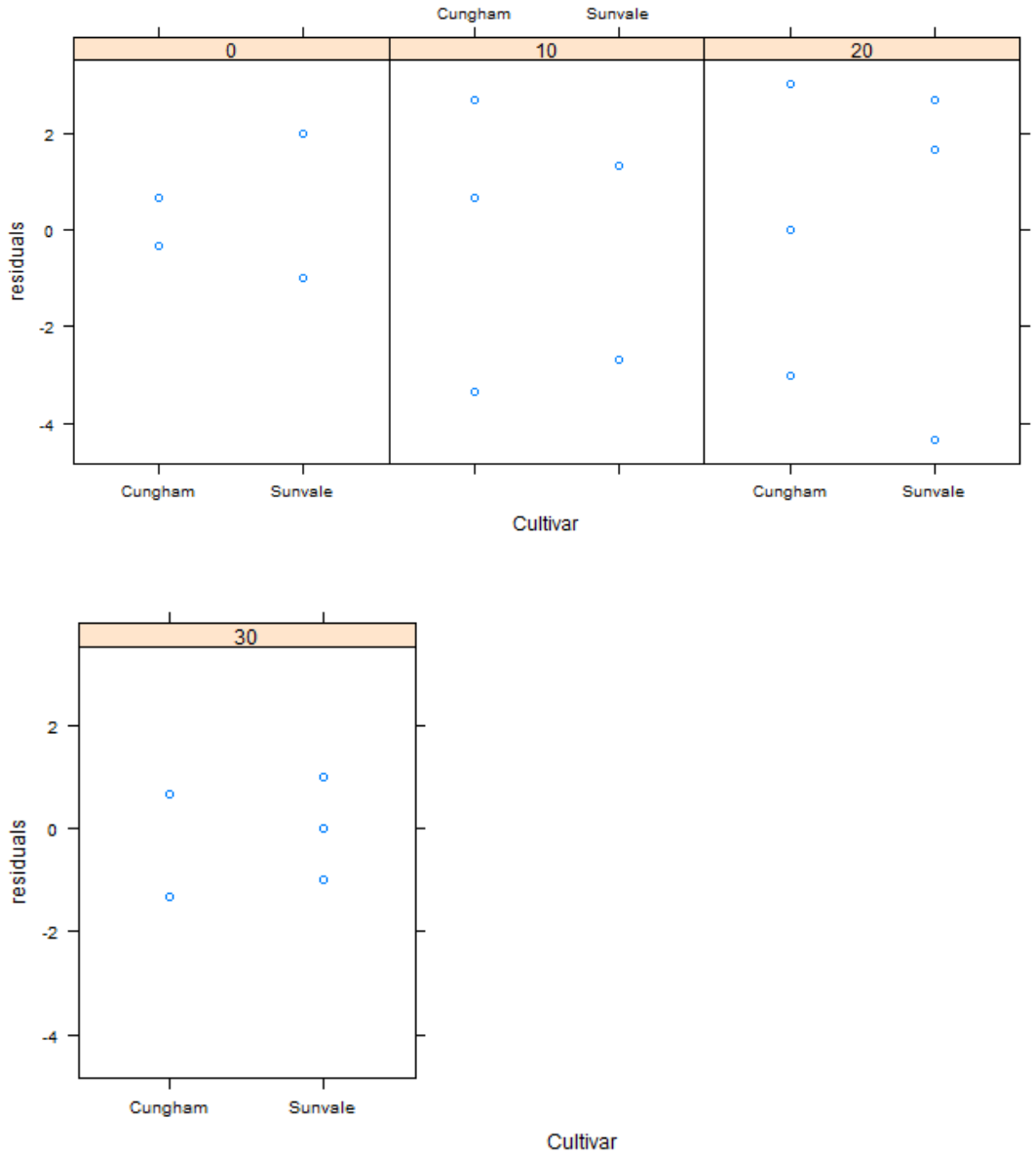

Figure 26 Plots of the residuals of the Cultivars per Salinity level

```
###We assess Normality in the usual manner, with a Normal probability plot
qqmath(Germ$residuals)
```

Figure 27 Syntax for plotting the normality of the residuals

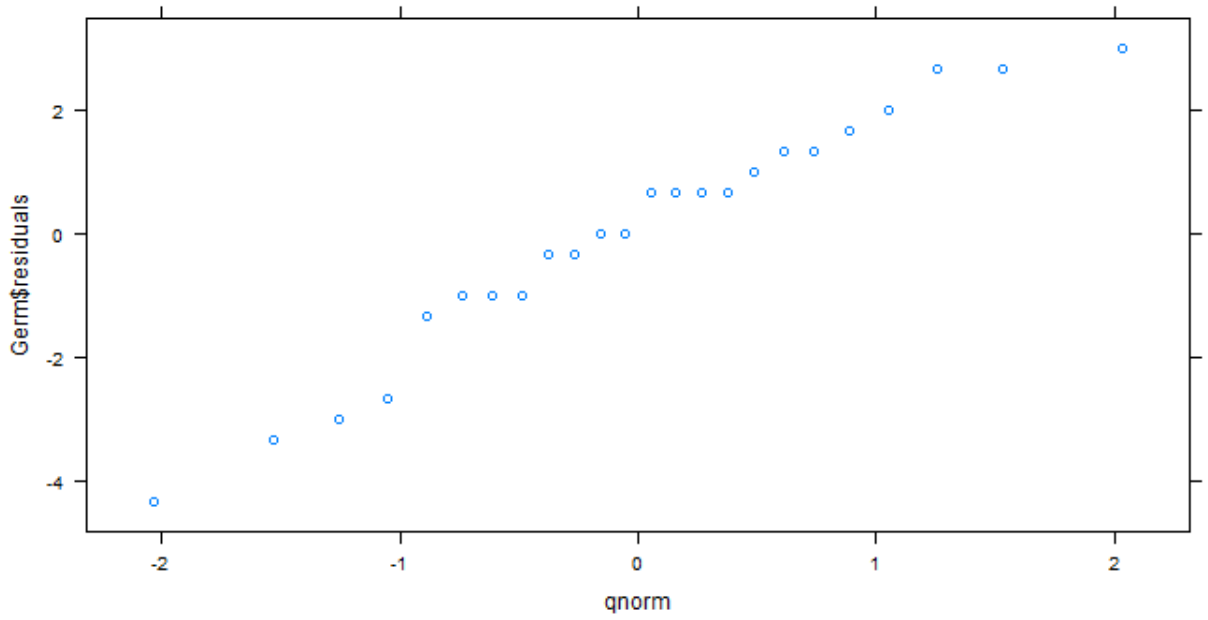

Figure 28 Normal probability plot

The closer this is to a straight line, with both bottom left and top right more or less fitting on the line indicate the data is probably normal. This is comforting as ANOVA requires that assumption to hold.

Results of ANOVA

The main effect of salinity was found to be significant. So the next step may be too look at which levels of salinity may be different. In the following code we show the use of a multiple comparison test. The output is explained below.

```
###Tukey's Honest Significant Difference for Salinity treatment ##difference

TukeyHSD(germ.aov, "Salinity", ordered = FALSE, conf.level = 0.95)
```

Figure 29 Checking for the significantly different Salinity levels

```
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Germination ~ Cultivar + Salinity + Cultivar * Salinity,
data = Germ)

$Salinity
             diff         lwr          upr     p adj
10-0   -4.666667  -8.525804  -0.8075290 0.0153399
20-0  -10.500000 -14.359138  -6.6408624 0.0000043
30-0  -14.000000 -17.859138 -10.1408624 0.0000001
20-10  -5.833333  -9.692471  -1.9741957 0.0026471
30-10  -9.333333 -13.192471  -5.4741957 0.0000187
30-20  -3.500000  -7.359138   0.3591376 0.0824904
```

Figure 30 The table for the significantly different Salinity levels

In the above output the first column show the comparison. So the first line show 10-0, which will be the result for the comparison of the 10% salinity with no salinity. the next column is the difference, between the means. The lwr and upr are the lower and upper confidence interval of this difference. The final column indicates the p value for the difference.

It is only appropriate to look at the differences if the F test in the ANOVA were significant. the Tukey's test above allow us to determine in this experiment which salinity levels are different. The note of `95% family-wise confidence level,` indicates that we are not exceeding an overall significance level of 95% in this experiment overall.

If you wish to ask more about any aspects in his demonstration, please post a question.