Biology Statistics using Genstat

There is a variety of statistical analysis you can carry out. You need to think about ***how*** you will do the analysis as you plan the experiment so you have recorded the correct amount of data.

- You could do a t-test but this only compares 2 treatments and you will probably have more treatments in your investigation.
- Regression analysis allows you to find out if there is a relationship between variables but again only looks an explanatory and a response variable and you will probably have more variables
- Anova or analysis of variance will tell you if at least one of the treatment means is significantly different from the rest. Further investigation may be required.
- Box and Whisker graphs (especially with the error bars) will allow you to see which differences are statistically significant
- Chi-squared tests will allow you to test a hypothesis to see whether there is an association between two variables. Allows you to test for no association (null hypothesis) vs an association It is useful however if your data is discrete – e.g. you have counted the number of something or have frequencies

Genstat Programs

There are actually two programs, the analysis program where your spreadsheet is and the graphing program. You can access these by clicking on the icons on the task bar. for the graphs and for the spreadsheet and analysis.

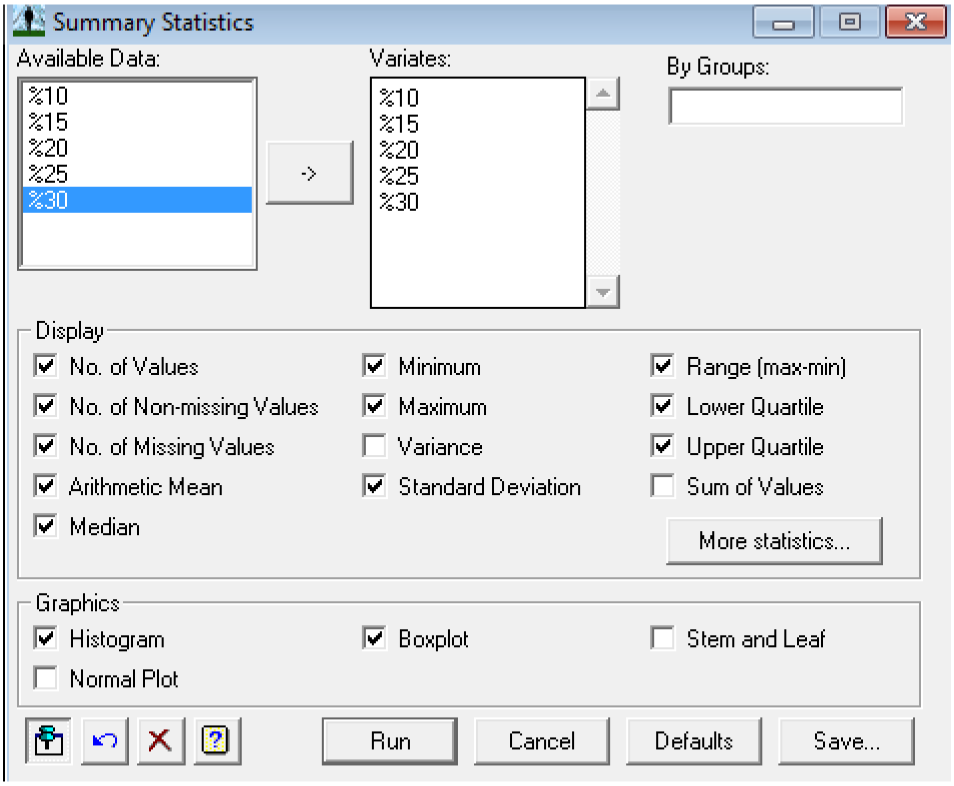
Summary statistics

When you wish to find the mean, median, standard deviation etc.

Open Genstat (its found on Masters/Maths/Maths with stats/Genschool2011)

You may copy this folder onto a memory stick if you like.

1. Open Genstat GenStat for Teaching a...
2. Click on
3. Open the file ***Crabs*** (Browse and look in the biology folder next to the Genschool 2011 folder)
4. Click on **Summary Statistics** from the **Stats** menu along the top
5. Choose the data you want the statistics for and which graphs you want


Summary Statistics

Available Data:
%10
%15
%20
%25
%30

Variates:
%10
%15
%20
%25
%30

By Groups:

Display
☑ No. of Values
☑ No. of Non-missing Values
☑ No. of Missing Values
☑ Arithmetic Mean
☑ Median
☑ Minimum
☑ Maximum
□ Variance
☑ Standard Deviation
☑ Range (max-min)
☑ Lower Quartile
☑ Upper Quartile
□ Sum of Values

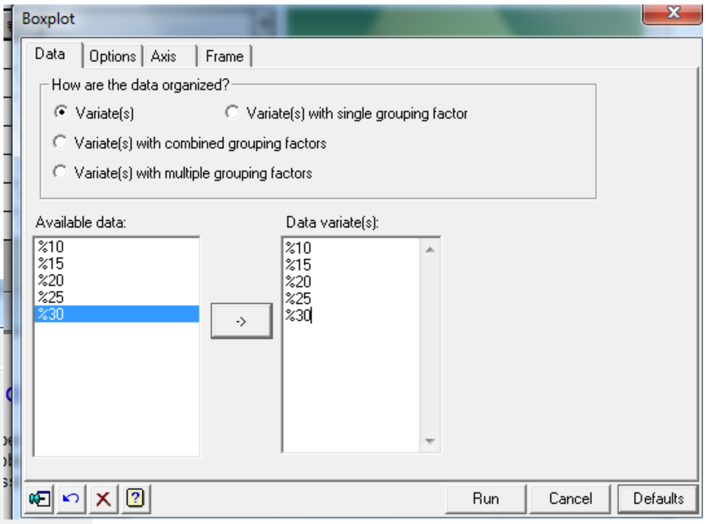
Graphics
☑ Histogram
□ Normal Plot
☑ Boxplot
□ Stem and Leaf

6. The graphs will show in the graphics window and the ⇦ ⇨ allow you to move from one graph to another. You can copy these graphs into word using right click or 📋

Graphs

To graph all the data on one boxplot choose **Boxplot** from the **Graphics** menu

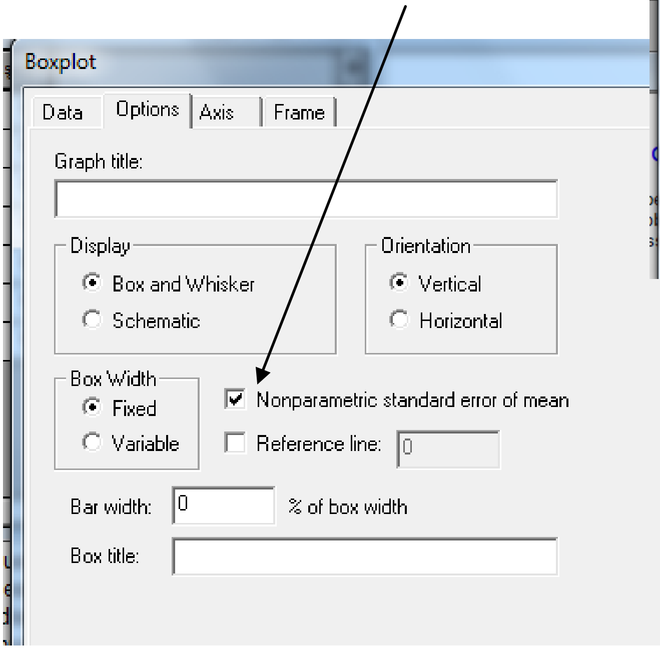
- Click all the variables across
- Click on options and tick


Boxplot

How are the data organized?
● Variate(s)
○ Variate(s) with single grouping factor
○ Variate(s) with combined grouping factors
○ Variate(s) with multiple grouping factors

Available data:
%10
%15
%20
%25
%30

Data variate(s):
%10
%15
%20
%25
%30


Boxplot

Graph title:

Display
● Box and Whisker
○ Schematic

Orientation
● Vertical
○ Horizontal

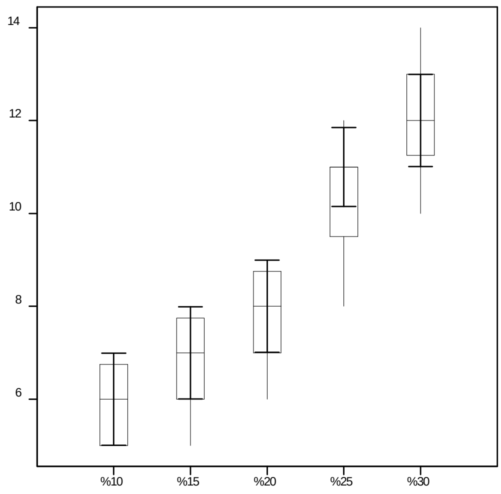
Box Width
● Fixed
○ Variable

☑ Nonparametric standard error of mean
□ Reference line: 0

Bar width: 0 % of box width

Box title:

This will produce a box plot with error bars. Note: If the error bars overlap there is no significant difference between the variates.



You can see the error bars on the two graphs on the right overlap so they are not significantly different.

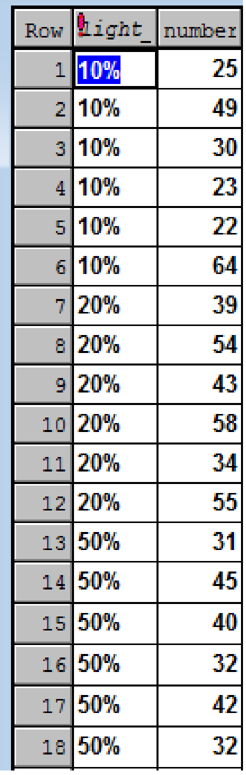
Anova

T his file has the data for light levels and radish hypocotysis. There are four treatments – the 4 different light levels, so you cannot use a t-test.

This is easily done in Genstat but you must have 2 columns, one for the treatment (factor) and one for the recorded data. Just type (or open your excel sheet) in Genstat.

Then right click on the column for the treatment and choose **Convert to Factor**

Your sheet should look like the one opposite



Row	light_	number
1	10%	25
2	10%	49
3	10%	30
4	10%	23
5	10%	22
6	10%	64
7	20%	39
8	20%	54
9	20%	43
10	20%	58
11	20%	34
12	20%	55
13	50%	31
14	50%	45
15	50%	40
16	50%	32
17	50%	42
18	50%	32

Now just choose **Anova** from the **Stats** menu

You will get the following Output (select **Output** from the **Window** menu)

Analysis of variance

Variate: number

```
Source of variation          d.f.        s.s.        m.s.      v.r.    F pr.
light_levels                    3      1155.7       385.2      3.39    0.038
Residual                       20      2275.7       113.8
Total                          23      3431.3
```

<0.05 so statistically significant

Information summary

All terms orthogonal, none aliased.

Message: the following units have large residuals.

```
*units* 6                                      28.5          s.e.   9.7
```

Outlier in row 6 of your spreadsheet

Tables of means

Variate: number

Grand mean 36.8

```
     light_levels       10%       100%        20%        50%
                       35.5       27.7       47.2       37.0
```

Standard errors of means

```
Table              light_levels
rep.                          6
d.f.                         20
e.s.e.                     4.35
```

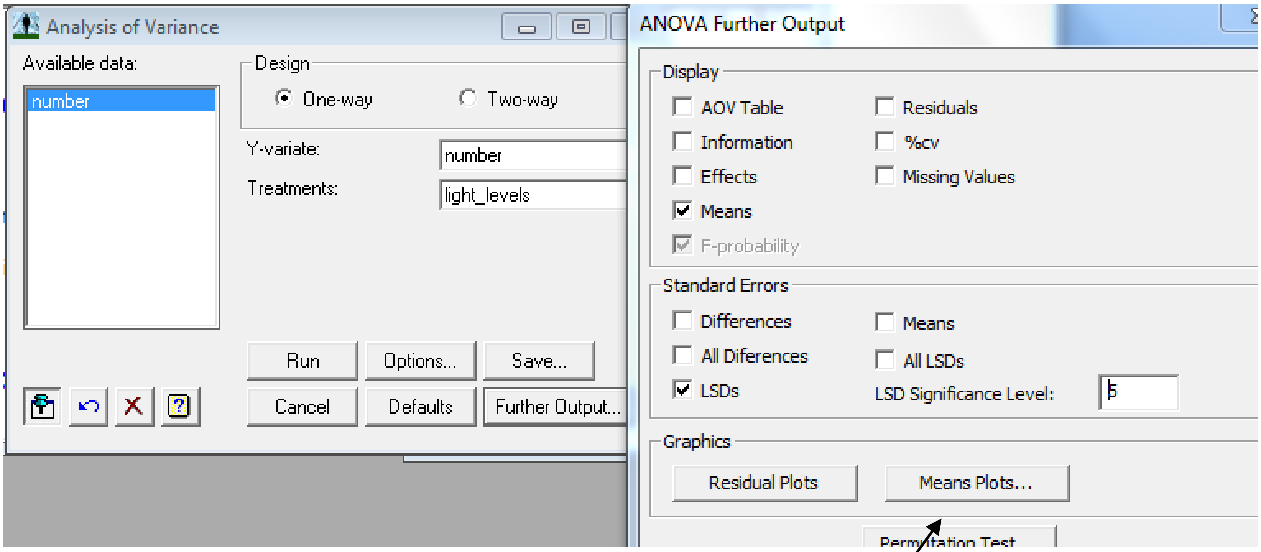
6 replicates

20 degrees of freedom

4 treatments *(6-1) replicates

Standard errors of differences of means

```
Table              light_levels
rep.                          6
d.f.                         20
s.e.d.    6.16
```



Analysis of Variance

Available data:
number

Design
● One-way ○ Two-way

Y-variate: number

Treatments: light_levels

Run Options... Save...

Cancel Defaults Further Output...

ANOVA Further Output

Display

□ AOV Table □ Residuals

□ Information □ %cv

□ Effects □ Missing Values

☑ Means

☑ F-probability

Standard Errors

□ Differences □ Means

□ All Diferences □ All LSDs

☑ LSDs LSD Significance Level: 5

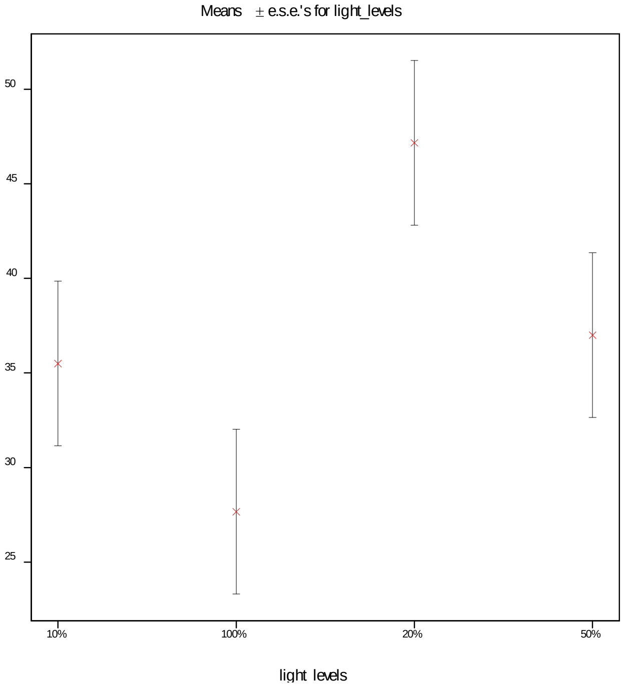
Graphics

Residual Plots Means Plots...

Permutation Test

There are a number of graphs you could get (look at further output) but

Gives you an informative graph... there should be no overlap between the bars if there is a difference.

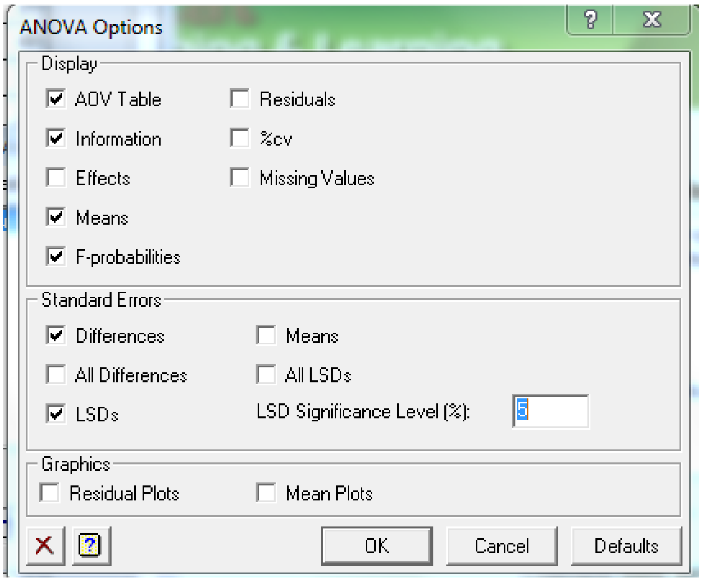


So we can see that there is no significant difference between 10%, 100% and 50% but there appear to be a significant difference between these three and 20%

You can also use the ***least significant difference*** to check there is a significant difference in any two pairs of treatments.

To do this

- Click on LSD in ANOVA options as shown on the next page
- The 5 % level will give you a 95% confidence level



ANOVA Options

Display
☑ AOV Table ☐ Residuals
☑ Information ☐ %cv
☐ Effects ☐ Missing Values
☑ Means
☑ F-probabilities

Standard Errors
☑ Differences ☐ Means
☐ All Differences ☐ All LSDs
☑ LSDs LSD Significance Level (%): 5

Graphics
☐ Residual Plots ☐ Mean Plots

You get the following in the output

```
Least significant differences of means
(5% level)

Table                    light_levels
rep.                                6
d.f.                               20
l.s.d.                          12.85
```

You now use the lsd of 12.85 to compare a pair of treatments

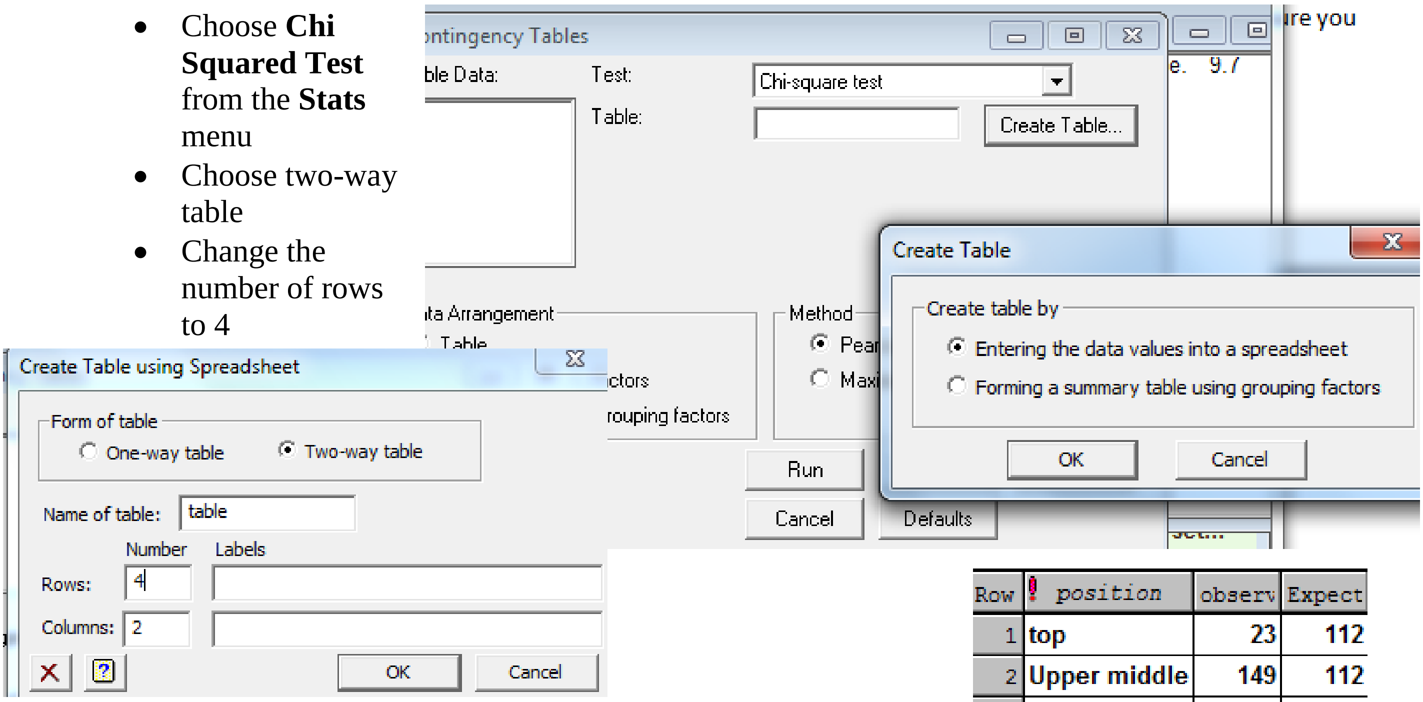
Diff between 10 + 20% = 47.2-35.5 = 11.7 <12.87 so no significant difference at the 95% confidence level
Diff between 10 + 50% = 37-35.5 =1.5 <12.87 so no significant difference at the 95% confidence level
Diff between 10 + 100% = 27.7-35.5 = -7.8 <12.87 so no significant difference at the 95% confidence level
Diff between 20 + 50% = 47.2-37 = 10.2<10.2 so no significant difference at the 95% confidence level
Diff between 20 + 100% = 47.2 – 27.7 = 19.5 >12.87 so there is a significant difference at the 95% confidence level
Diff between 50 + 100% =37 – 27.7 = 9.3 <12.87 so no significant difference at the 95% confidence level

Chi-Squared test

This data is the distribution of mangrove Pneumatophores (breathing roots that appear above the ground). the null hypothesis is no association
between the number of pneumatophores and the level on the beach versus the alternative hypothesis of an association between. the numbers of pneumatophores at the four levels on the beach that she investigated.

For this you need to enter in your observed and expected values in a table in Genstat.

- Choose **Chi Squared Test** from the **Stats** menu
- Choose two-way table
- Change the number of rows to 4

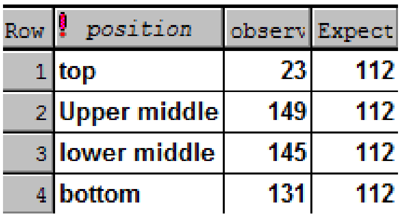


ontingency Tables
ble Data:
Test: Chi-square test
Table:
ta Arrangement
Table
ctors
rouping factors
Method
● Pear
○ Maxi

Create Table
Create table by
● Entering the data values into a spreadsheet
○ Forming a summary table using grouping factors

Create Table using Spreadsheet
Form of table
○ One-way table ● Two-way table
Name of table: table
Number Labels
Rows: 4
Columns: 2

- Enter in your data – either the means or the totals



Row	position	observ	Expect
1	top	23	112
2	Upper middle	149	112
3	lower middle	145	112
4	bottom	131	112

In the table above I have edited the first column but you don't need to do this as all you are looking for is whether there is a significant difference between any of the treatments

- Now look at the **Output** from the **Window** menu

Chi-square test for association between C16 and C17

Likelihood chi-square value is 74.90 with 3 d.f.

Probability level (under null hypothesis) p < 0.001

Note: 3 df as 4 treatments

Note: the p value is <0.001 so it is very highly significant statistically

(p<0.05 is significant, p<0.01 highly significant and p<0.001 so it is very highly significant statistically)

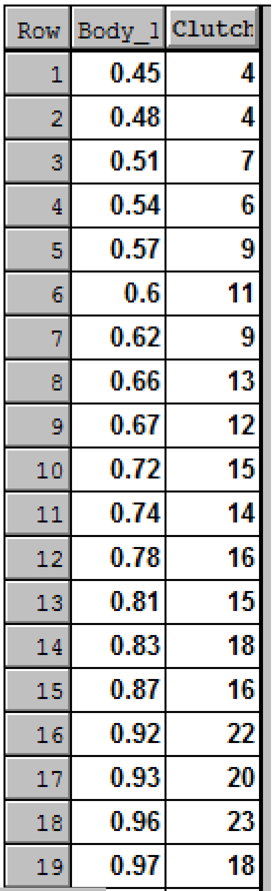
Linear Regression

This is used when you want to find a relationship between two variables. One the independent or explanatory variable is used to predict the dependent or response variable.

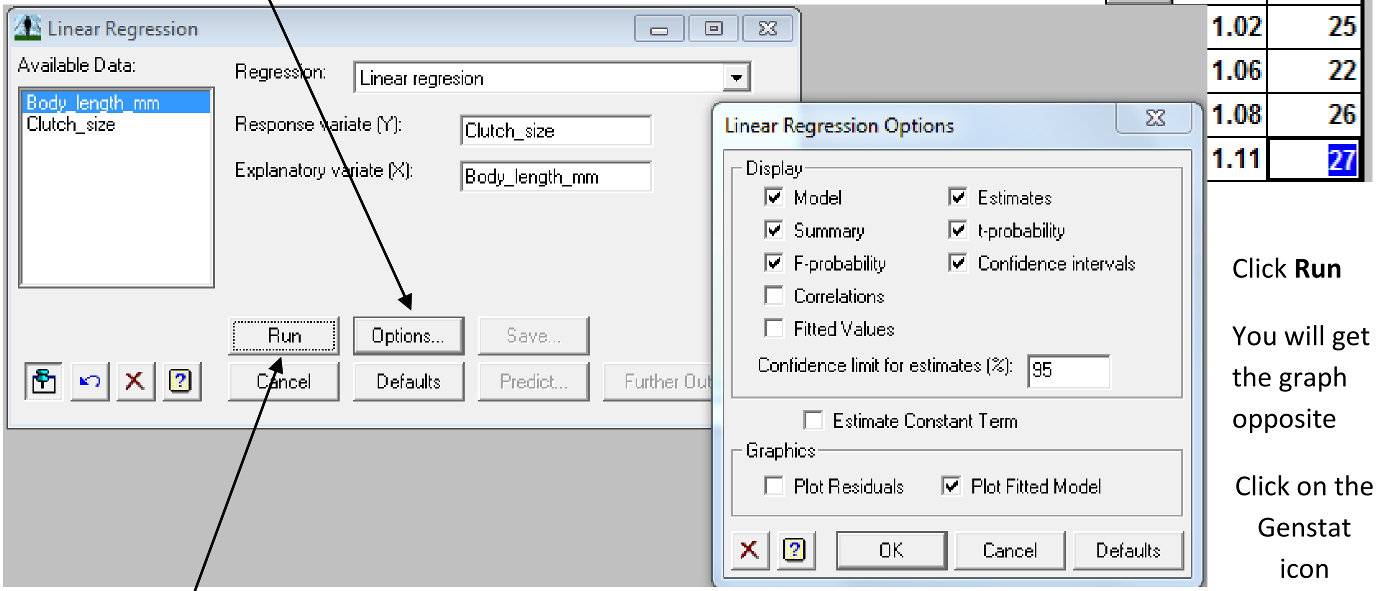
The data file Clutch size opposite, looks at the body length in mm and the clutch size (the number of eggs carried) in a small freshwater crustacean called Daphnia.

To perform Linear regression, you simply choose **Linear Regression** from the **Stats** menu

Click on **options** to get further output. Notice we are trying to predict the clutch size from the body length.



Row	Body_l	Clutch
1	0.45	4
2	0.48	4
3	0.51	7
4	0.54	6
5	0.57	9
6	0.6	11
7	0.62	9
8	0.66	13
9	0.67	12
10	0.72	15
11	0.74	14
12	0.78	16
13	0.81	15
14	0.83	18
15	0.87	16
16	0.92	22
17	0.93	20
18	0.96	23
19	0.97	18
	1.02	25
	1.06	22
	1.08	26
	1.11	27



Linear Regression

Available Data: Body_length_mm, Clutch_size

Regression: Linear regresion

Response variate (Y): Clutch_size

Explanatory variate (X): Body_length_mm

Linear Regression Options

Display

☑ Model ☑ Estimates

☑ Summary ☑ t-probability

☑ F-probability ☑ Confidence intervals

☐ Correlations

☐ Fitted Values

Confidence limit for estimates (%): 95

☐ Estimate Constant Term

Graphics

☐ Plot Residuals ☑ Plot Fitted Model

Click **Run**

You will get the graph opposite

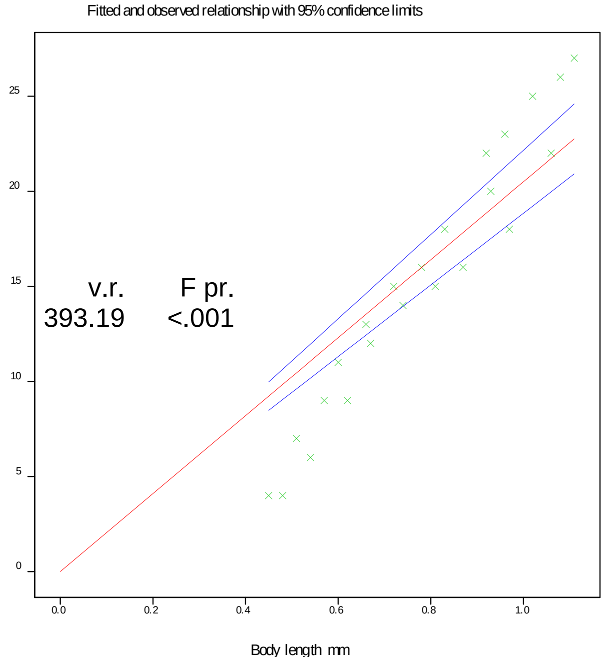
Click on the Genstat icon

and look under the **Window** menu for **output**

The blue lines on the graph are the confidence interval lines. As we have used a sample from a population to estimate, the95% confidence interval means that there is a 95% probability that the population results will lie inside the interval.

Regression analysis

Response variate: Clutch_size
Fitted terms: Constant, Body_length_mm



Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	1005.18	1005.184	393.19	<.001
Residual	21	53.69	2.556		
Total	22	1058.87	48.130		

Percentage variance accounted for 94.7
Standard error of observations is estimated to be 1.60.

Message: the following units have large standardized residuals.

Unit	Response	Residual
19	18.00	-2.38

Estimates of parameters

Parameter	estimate	s.e.	t(21)	t pr.
Constant	-10.42	1.34	-7.78	<.001
Body_length_mm	33.05	1.67	19.83	<.001

Parameter	lower95%	upper95%
Constant	-13.20	-7.633
Body_length_mm	29.59	36.52

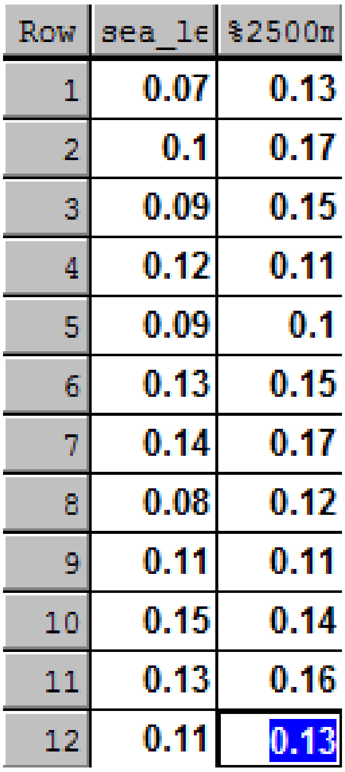
This tells you there is a strong relationship, (notice the t probability is <0.01) and that body length is highly significant in predicting the clutch size.

The exact relationship is that the clutch size = 33.05 times the body length – 10.42

Notice also that 94.7% of the variation in the clutch size can be accounted for by the linear regression model that has been fitted (this is the R^2 value).

Paired t-test

This data is the effect of alcohol consumption on the body which appears to be greater at higher altitude. 12 subjects consume 100cc of alcohol in a drink at 2500m and the amount of alcohol in their blood is tested 2 hours later. Two weeks later the same subjects are tested at sea level. The null hypothesis is that there is no difference between the means at sea level and 2500m



Row	sea_le	%2500m
1	0.07	0.13
2	0.1	0.17
3	0.09	0.15
4	0.12	0.11
5	0.09	0.1
6	0.13	0.15
7	0.14	0.17
8	0.08	0.12
9	0.11	0.11
10	0.15	0.14
11	0.13	0.16
12	0.11	0.13

One-sample t-test

Variate: Y[1].

Summary

Sample	Size	Mean	Variance	Standard deviation	Standard error of mean
%2500m-sea_level	12	0.02667	0.0007333	0.02708	0.007817

95% confidence interval for mean: (0.009461, 0.04387)

Test of null hypothesis that mean of %2500m-

Notice zero is not in the confidence interval, so at the 95 %level there is a difference

Test statistic t = 3.41 on 11 d.f.

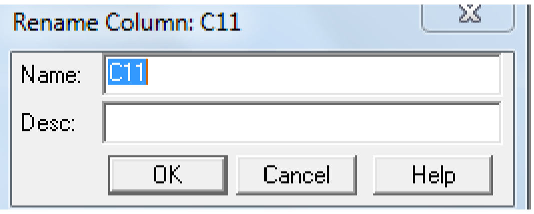
Probability = 0.006

Note the probability that there is a difference is 0.006 – so the Null hypothesis should be rejected.

Doing your own analysis on your own data!

This is straightforward…

1. Look carefully how the example files are laid out. Use the same type of layout.
2. Click on **New** from the **File** menu and click on the **Spreadsheet icon**
3. Decide how many rows and columns you need (though icons let you delete or insert rows and columns)
4. Type in your data – you can label the columns by placing the cursor at the frontof the cell until a pencil appears and then clicking
5. You can save as a genstat file (.gsh) or an Excel file (.xls)
6. Or alternatively set up your data in Excel, checking that you have no extra empty rows or columns and then just open it in Excel.



Rename Column: C11

Name: C11

Desc:

Good luck!!!!

Data Files used with permission from

Biozone, Level 8 Biology a student Workbook ISBN 9789-1-877462-50-4

ESA Publications, Year 13 Biology Learning Workbook, G Moriarty, A Roberts, M Sinclair, ISBN 978-1-877459-90-0

Continuing Biology, M Bayley & D Layzell ISBN 978-0-582-54783-4