# Appendix 2: Examples of Analyses in SPSS

**Survey of Deer Pellets in the Takahe Special Area, Murchison Mountains**
**MANIPULATING DATA IN SPSS**
**COMPARISON OF MEANS**
**ONE FACTOR ANOVA**

**Excel data file: TAKAHE.xls (DME CHCRO-28738)**

Data is from multiple lines within four catchments of the Takahe special area. The catchments are Chester Burn, Ettrick Burn, Snag Burn and Point Burn. Lines extended from the valley floor and up the side of the valley. Along each line at least 10 plots placed 15m apart. Each plot was 1.14m in radius and the presence and absence of deer pellets were noted. A larger plot, 2.5m in radius was also measured at the same place and the number of groups of deer pellets counted. Some of the lines were longer than others. These longer lines had more than 10 plots on a line and extended up to the treeline (or thereabouts). The variable **type** is set as 1 for plots on the "short" section of the line in the valley bottom, or 2 for any plots that are beyond the valley bottom and on the "long" section of the line.

First the dataset **Takahe.xls** is separated into two separate files, one for the short lines and one for the data from the long lines.

Open the Excel spreadsheet file Takahe.xls (DME CHCRO-28738)

1. Copy and paste the data into SPSS Data editor window. Go into variable view (see bottom of screen), under name insert appropriate variable names e.g., label the first variable **line**.
2. Click on **Data** and then **Select Cases**
3. **Select** the variable Type and use an **If** condition such If **type** = 1. Set **Unselected cases are Deleted**. **Save** this new file as **Tak_short**.
4. Repeat the above step for the type 2 data and **Save** this new file as **Tak_long**.

***Comparison among catchments for the lower sections of the lines***

When the data for the short lines is summarised it is clear that most of the plots did not have any deer pellets. For example, at Chester Burn there were 497 plots over 50 short-lines. Of these 0.9054% had some pellets in the 1.14m radius circle, and within the 2.5m radius plots the average number of groups was 0.1187.

*Descriptive Statistics - Chester Burn*

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| GROUPS | 497 | .00 | 3.00 | .1187 | .4016 |
| PELLETS | 497 | .00 | 1.00 | 9.054E-02 | .2872 |

*Descriptive Statistics - Ettrick Burn*

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| GROUPS | 493 | .00 | 5.00 | .1034 | .4722 |
| PELLETS | 493 | .00 | 1.00 | 2.434E-02 | .1543 |

*Descriptive Statistics - Snag Burn*

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| GROUPS | 496 | .00 | 5.00 | .1250 | .4762 |
| PELLETS | 496 | .00 | 1.00 | 3.427E-02 | .1821 |

*Descriptive Statistics - Point Burn*

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| GROUPS | 495 | .00 | 6.00 | 8.485E-02 | .4554 |
| PELLETS | 495 | .00 | 1.00 | 2.424E-02 | .1540 |

Data that has so many zeroes can be difficult to analyse. In this case with so few data points that did not equal zero it would be difficult to detect any trends or patterns. The plots within lines can be aggregated to create a new variable: the total number of pellet groups on the plots and the number of plots where pellets were present within each line. Because there were the same number of plots (10) within each short-line the totals can be used to compare among lines. If there were different numbers of plots on the lines averages should be used. Using either file, e.g., **tak_short.sav**

1. Click on **Data** and then **Aggregate**
2. The Break variable is **catchment** and line and the **Aggregate** variables are **groups** and **pellets**. Change the function of the default aggregating from the mean to the **sum** by clicking on **function** and then selecting **sum of values**. Rename the aggregate variables as **sumgrps** and **sumplts**.
3. **Save** the aggregated data into a new file, e.g., **tak_short_aggr.sav**.

The total number of pellet groups and total number of plots with pellets present can be summarised by each catchment using the file **tak_short_aggr.sav**.

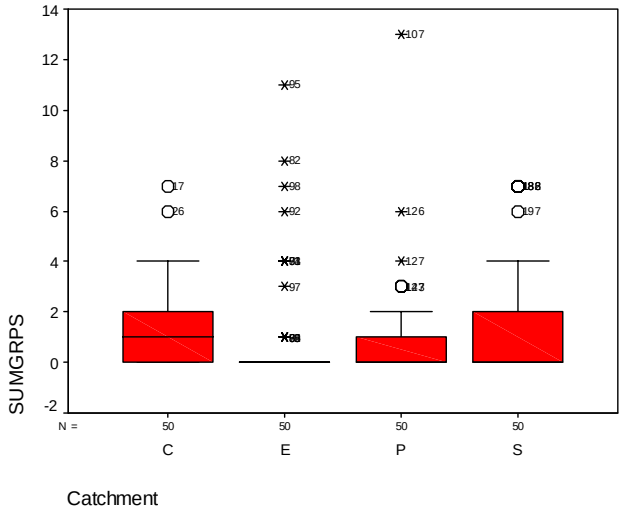
1. To produce seperate summaries for each catchment first click on **Data** and then **Split File**.
2. In the Split File dialogue box **Create Groups** based on **catchment** and select **Compare Groups**.
3. Click on **Analyse**, then **Descriptive statistics** and then **Descriptives**. The variables to analyse are **sumgrps** and **sumplts**.

*Descriptive Statistics*

| Catchment | | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|---|
| C | SUMGRPS | 50 | .00 | 7.00 | 1.1800 | 1.5477 |
| | SUMPLTS | 50 | .00 | 4.00 | .9000 | 1.1294 |
| E | SUMGRPS | 50 | .00 | 11.00 | 1.0200 | 2.3861 |
| | SUMPLTS | 50 | .00 | 4.00 | .2400 | .6869 |
| P | SUMGRPS | 50 | .00 | 13.00 | .8400 | 2.1415 |
| | SUMPLTS | 50 | .00 | 3.00 | .2400 | .5911 |
| S | SUMGRPS | 50 | .00 | 7.00 | 1.2400 | 1.9437 |
| | SUMPLTS | 50 | .00 | 3.00 | .3400 | .7174 |

The total number of pellet groups along each line within catchments can be displayed graphically using boxplots. Remember to turn off the **Split File** option in **Data** before you create the boxplots.

1. Click on **Graphs** and then **Boxplot**.
2. In the Boxplot Dialogue box click on **Simple**. The **Data in Charts** are **Summaries for groups of cases**.
3. Click on **Define** and select **sumgrps** as the **Variable** that the boxes represent. The **Category Axis** is **catchment**.
4. Click on **Continue** and then **OK**.



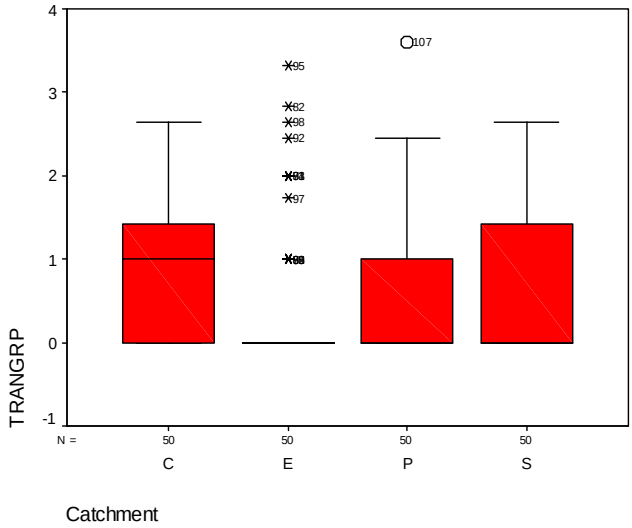

It appears that the catchment Ettrick Burn has mostly low counts of pellet groups along lines, although a few lines do have some pellet groups. The catchments all have about the same spread of pellet group counts, i.e., all catchments have lines with no pellet groups while, at Point Burn, lines have up to 13 pellet groups.

To test if there are signficant differences among the catchments ANOVA can be used. However the data appears to be very non-normally distributed, that is, there are many lines with few, if any pellet groups and a few with many pellet groups. A transformation may improve this. For example, a square root transformation can be used.

1. Click on **Transform** and then **Compute**.
2. In the Compute dialogue box the **Numeric Expression** is **sqrt(sumgrps)**. The **Target Variable** is the new transformed variable, e.g., **trangrp**.
3. Click on **OK**

The transformation has improved the normality slightly but the distribution is still very skewed (to the right). There are other transformations that can be used such as taking the natural log of the raw data. This transformation and others that involve dividing by the raw data (e.g., reciprocal) have problems when there are zero's in the data.

Boxplots of the transformed data can be created using the same procedure as above.




Given the assumption of normality for ANOVA is the least serious assumption to violate analysis of variance can be conducted on the transformed numbers of pellet groups.

1. Click on **Analyse**, then **General Linear Models** and then **univariate**.
2. In the GLM - **univariate** dialogue box the **Dependent Variable** is trangrp and the **Fixed Factor** is **catchment**.
3. Click on **OK.**

**Tests of Between-Subjects Effects**

**Dependent Variable: TRANGRP**

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 4.825 | 3 | 1.608 | 2.316 | .077 |
| Intercept | 73.076 | 1 | 73.076 | 105.239 | .000 |
| Catchment | 4.825 | 3 | 1.608 | 2.316 | .077 |
| Error | 136.099 | 196 | .694 | | |
| Total | 214.000 | 200 | | | |
| Corrected Total | 140.924 | 199 | | | |
| Corrected Total | 140.924 | 199 | | | |

a R Squared = .034 (Adjusted R Squared = .019)

There is no evidence that there are significant differences among catchments, ($F_{3, 196}$ = 2.316 and P = 0.077) if $\alpha$ = 0.05 is used. If there were evidence the ANOVA could be repeated with a post-hoc test such as Least Significant Difference (LSD) to see whether the catchment differences were. For the sake of the example, $\alpha$ = 0.10 is used. A post hoc test can be conducted.

1. Click on **Analyse**, then **General Linear Models** and then **univariate**.
2. In the **univariate** dialogue box the **Dependent Variable** is trangrp and the **Fixed Factor** is **catchment** as before.
3. Click on **Post Hoc** and in the dialogue box the **Post Hoc Test** is for **catchment**.
4. Select **LSD**

**Multiple Comparisons**

**Dependent Variable: TRANGRP - LSD**

| | | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| (I) Catchment | (J) Catchment | | | | Lower Bound | Upper Bound |
| C | E | .3351 | .167 | .046 | 6.413E-03 | .6638 |
| | P | .3510 | .167 | .036 | 2.234E-02 | .6797 |
| | S | 7.417E-02 | .167 | .657 | -.2545 | .4028 |
| E | C | -.3351 | .167 | .046 | -.6638 | -6.4130E-03 |
| | P | 1.593E-02 | .167 | .924 | -.3127 | .3446 |
| | S | -.2609 | .167 | .119 | -.5896 | 6.776E-02 |
| P | C | -.3510 | .167 | .036 | -.6797 | -2.2340E-02 |
| | E | -1.5927E-02 | .167 | .924 | -.3446 | .3127 |
| | S | -.2768 | .167 | .098 | -.6055 | 5.183E-02 |
| | S | -.2768 | .167 | .098 | -.6055 | 5.183E-02 |
| S | C | -7.4170E-02 | .167 | .657 | -.4028 | .2545 |
| S | C | -7.4170E-02 | .167 | .657 | -.4028 | .2545 |
| | E | .2609 | .167 | .119 | -6.7757E-02 | .5896 |
| | E | .2609 | .167 | .119 | -6.7757E-02 | .5896 |
| | P | .2768 | .167 | .098 | -5.1830E-02 | .6055 |
| | P | .2768 | .167 | .098 | -5.1830E-02 | .6055 |

The results can be confusing from such tests. By looking at the raw significance levels and using $\alpha$= 0.1 there is evidence that Chester Burn has significantly more pellet groups than Ettrick and Point Burn. Point Burn has significantly less pellet groups than Chester and Snag Burn. With a more

conservative significance level to allow for the multiple comparisons none of the catchments have a significant differences.

**Comparison among upper and lower sections of the line**

Consider the lines that extended up from the base of the valley to the upper slopes. The data from the plots beyond the initial 10 in the valley floor is in **tak_long.sav**. First the data needs to be aggregated along lines. This time there are different numbers of plots along a line so the average number of pellet groups in a line is used.

1. Click on **Data** and then **Aggregate.**
2. The **Break** variable is **catchment** and line and the **Aggregate** variables are **groups**. Keep the **function** at the default to calculate the mean. Rename the aggregate variables as **mgrpup** to refer to the mean number of groups of pellets on the upper part of the lines. **Save** the aggregated data into a new file, e.g., **tak_long_aggr**.

There were only 20 long lines - in Chester Burn the lines were number 11, 21, 31, 41, and 50. In Ettrick Burn the lines were number 4, 11, 20, 28 and 46. In Point Burn they were line number 12, 21, 30, 39 and 50. In Snag Burn they were line number 7, 16, 24, 37, and 50. Because these 20 lines passed through both the valley floor and the valley sides the data can be used to compare the number of pellet groups by a paired t-test.

First a new file is created which combines the line averages for the number of pellet groups for the valley floor section of the line (the first 10 plots) and the valley side section of the line (the plots numbered 11 onwards). To do this a new variable can be added into the existing file **tak_long_aggr**. The new variable is the mean number of groups of pellets on the lower, valley floor section of the line. This data needs to be aggregated from the file **tak_short.sav**. Open this file.

1. Click on **Data** and then **Aggregate**
2. The **Break** variable is **catchment** and line and the **Aggregate** variables are **groups**. Keep the **function** at the default to calculate the mean. Rename the aggregate variables as **mgrplow** to refer to the mean number of groups of pellets on the lower part of the lines. Rather than saving this file click on the option **Replace working data file**.
3. The next step is to delete all the lines that do not extend to the upper slopes, i.e., delete all but 20 lines using **Edit** and **Clear**.
4. The 20 remaining lines should now correspond to the 20 lines that are in tak_long_aggr. Therefore the column that has the variable **mgrplow** can be copied (**Edit, Copy**) and pasted into **tak_long_aggr** (**Edit, Paste**).

The t-test for the difference between the upper and lower sections of the line can be computed.

1. Click on **Analyse**, then **Compare Means** and then on **Paired Sample T-Test**.
2. Select **mgrplow** and **mgrpup** as the **Paired Variables** for the test.
3. Click on **OK**.

**Paired Samples Test**

| | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|
| | | | | Lower | Upper | | | |
| MGRPLOW - MGRPUP | -1.2519E-02 | 7.236E-02 | 1.618E-02 | -4.6387E-02 | 2.135E-02 | -.774 | 19 | .449 |

There is no evidence of a difference in the average number of pellets from the upper and lower sections of the lines ($t_{19}$ = -0.774, P = 0.449). The average number of groups of pellets in plots in the lower section of the line was 0.03000 ± 0.01469. The average number of groups of pellets in plots in the upper section of the line was 0.04252 ± 0.01238. These summary statistics are part of the SPSS output.

While overall there may be no difference it would be interesting to see if there were differences among each catchment. This can be done by ANOVA of the difference between the upper and lower section for each line.

1. Create a new variable that is the difference between the average number of pellet groups on the upper and lower section. Click on **Transform** and **Compute**. In the **Compute dialogue** box the **Target Variable** can be named **diff** and the **Numeric Expression** is **mgrpup - mgrlow**.
2. Click on **OK**

In the ANOVA a post hoc test for **catchment** is calculated.

1. Click on **Analyse**, then **General Linear Models** and then **univariate**.
2. In the **univariate** dialogue box the **Dependent Variable** is diff and the **Fixed Factor** is **catchment** as before.
3. Click on **Post Hoc** and in the dialogue box the **Post Hoc Test** is for **catchment**.
4. Click on **Continue**
5. Click on **Options**
6. In the Options dialogue box select the variable **catchment** in the **Display Means For**.
7. Click on **Continue**
8. Click on **OK**

| | Mean | Std. Error | 95% Confidence Interval | |
|---|---|---|---|---|
| Catchment | | | Lower Bound | Upper Bound |
| C | 3.247E-02 | .025 | -1.971E-02 | 8.464E-02 |
| E | 2.964E-02 | .025 | -2.254E-02 | 8.182E-02 |
| P | 6.042E-02 | .025 | 8.243E-03 | .113 |
| S | -7.245E-02 | .025 | -.125 | -2.027E-02 |

The lines in Chester, Ettrick and Point Burn all had, on average, more pellet groups in the upper section of the line than in the lower sections while at Snag Burn there were more in the lower sections of the lines. The multiple comparisons of average pellet groups in Snag Burn compared with these three other catchments are all significant.

## Monitoring of Blue Cod from Paterson Inlet 1994 to 1998

### TWO FACTOR ANOVA

**Excel data file COD.xls (DME CHCRO-28733)**

Counts of blue cod and the length of each cod from 40 sites were recorded between 1994 and 1998. Twenty of the sites were inside the proposed reserve (Reserve = 1) and 20 were outside (Reserve = 0). The cod that were measured at each site over time are assumed to be different cod. That is, the same cod is not measured each year. If this assumption were not valid a *repeated measures analysis* would be needed.

To test whether there are differences in the average cod size (**mean**), the number of cod (**n**) and the variation in the cod sizes ANOVA is used. First a new variable, cv must be created. The cv is the standard deviation (**sd**) of the cod at each site divided by the average cod length at the site.

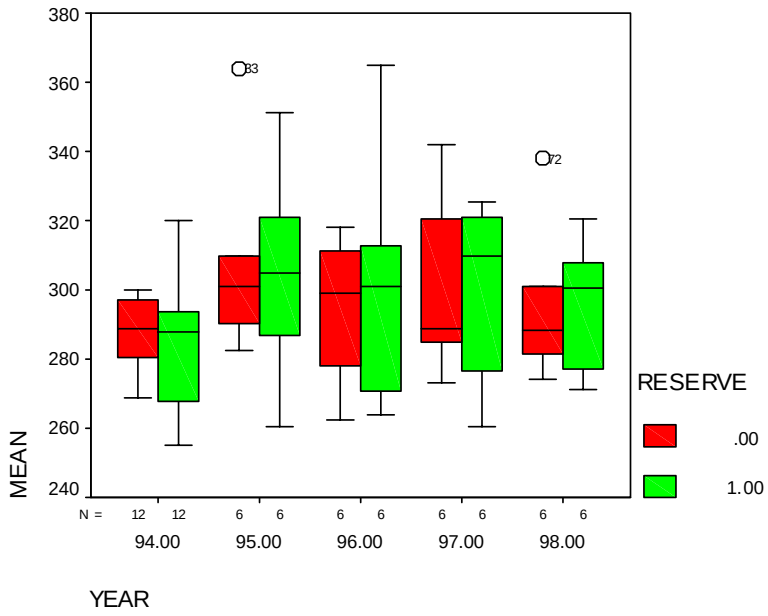
Open the Excel spreadsheet file Dataset COD.xls (DME CHCRO-28733)

1. Copy and paste the data into SPSS Data editor window. Go into variable view (see bottom of screen), under name insert appropriate variable name, e.g., label the first variable year.
1. Click on **Transform** and then **Compute**.
2. In the Compute dialogue box the **Numeric Expression** is sd/mean. The **Target Variable** is the new transformed variable, e.g., **cv**.
3. Click on **OK**

A two factor ANOVA is used. The factors are **reserve** and **year**. This will test for differences in cod within and outside the proposed reserve, for differences among the 5 years and the interactions of these factors.

Before ANOVA is used the data should be displayed e.g., by a boxplot or histogram to check that it is normally distributed and that there are homogeneous variances. To create boxplots for the variable **mean** for both levels of the variable **reserve** by **year**:

1. Click on **Graphs** and then **Boxplot**.
2. In the **Boxplot** dialogue box click on **Clustered**. The **Data in Charts** are **Summaries for Groups of Cases**.
3. Click on **Define** and select **mean** as the **Variable** and **year** as the **Category Axis** and **reserve** as the variable for **Define Clusters by**.
4. Click on **OK**.




There is no obvious trend in the data deviating from being normally distributed.

In the ANOVA the option for profile plots is selected. These plots are a useful way to display the data.

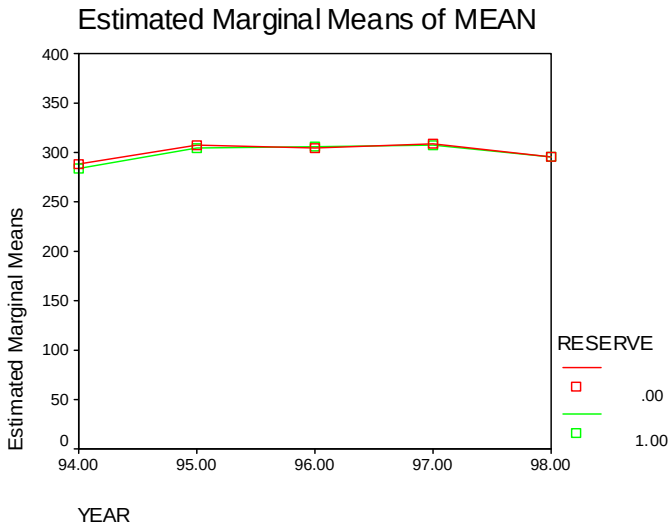
1. Click on **Analyse**, then **General Linear Models** and then **univariate**.
2. The **Dependent Variable** is mean and the **Fixed Factors** are **year** and **reserve**.
3. Click on **Model** and in the **univariate**: Model dialogue box select **Full Factorial**.
4. Click on **Continue**.
5. Click on **Plot** and in the **univariate**: **Profile Plots** dialogue box select **year** as the **Horizontal Axis** and **Separate Lines** for **reserve**.
   Click on **Add** so that the term **year*reserve** appears in the plots box.
6. Click on **Continue**.
7. Click on **OK**.

**Tests of Between-Subjects Effects**

**Dependent Variable: MEAN**

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 4263.872 | 9 | 473.764 | .842 | .581 |
| Intercept | 5893745.531 | 1 | 5893745.531 | 10468.747 | .000 |
| RESERVE | 3.118 | 1 | 3.118 | .006 | .941 |
| YEAR | 3953.086 | 4 | 988.272 | 1.755 | .149 |
| RESERVE * YEAR | 309.182 | 4 | 77.295 | .137 | .968 |
| Error | 34905.057 | 62 | 562.985 | | |
| Total | 6323967.347 | 72 | | | |
| Corrected Total | 39168.929 | 71 | | | |

a R Squared = .109 (Adjusted R Squared = -.021)




There is no evidence of a two-way interaction nor of any significant differences among levels of the main factors (P > 0.05 for all tests). The profile plot has two lines on it - one for the average cod size within the reserve and the other for the average size of cod in sites outside the reserve. The two lines are so close it is hard to see each separately. There does not appear to be any differences in the average cod length between inside, and outside the potential reserve sites. Further, this "lack of difference" was consistent among the years of the study.

The analysis can be repeated for the other dependent variables: **n** and **cv**.

**Monitoring of Vegetation at Whareorino Forest 1995 to 1999**

**REPEATED MEASURES ANALYSIS**

**Excel data file WHARE.xls (DME CHCRO-28739)**

Data from nine lines were collected between 1995 and 1999. Along each line there where a number of plots, between seven and 27 per line. Within the same plots each year the percentage foliage cover of trees was measured. This data represents *repeated measures* data since the same trees were measured.

Measurements of foliage cover from each tree within a plot can not be considered independent. We assume that the foliage cover in one plot is independent of the adjacent plot and therefore firstly the data is aggregated into plot averages.

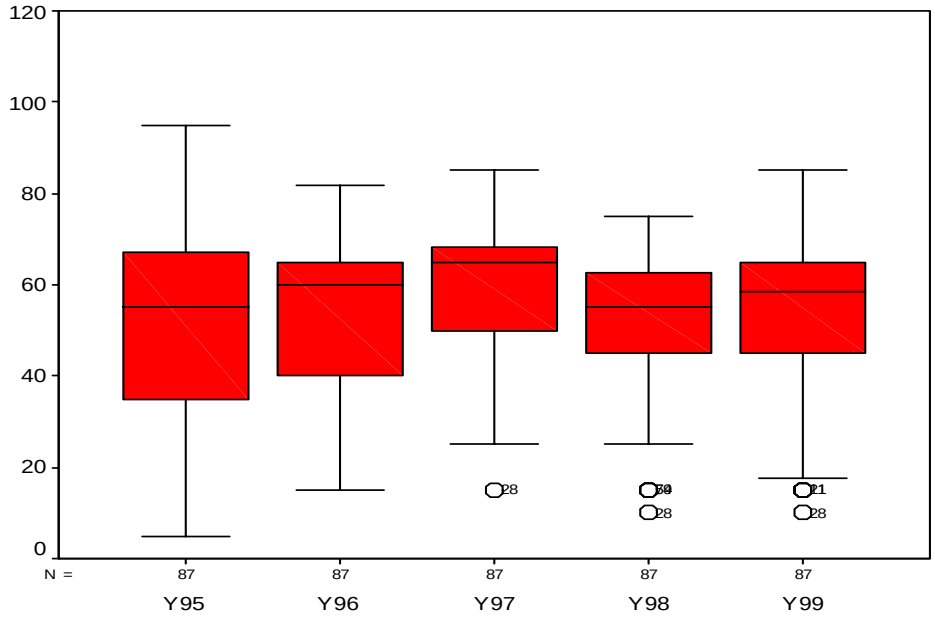
Open the Excel spreadsheet file Whare.xls (DME CHCRO-28739)

1. Copy and paste the data into SPSS Data editor window. Go into variable view (see bottom of screen), under name insert appropriate variable names, e.g., label the first variable line.
2. Click on **Data** and then **Aggregate**
3. The **Break** variables are **line** and **plot**, the **Aggregate** variables are **y95, y96, y97, y98, y99**. Rename these variables to remain as **y95, y96, y97, y98, y99**. **Save** the aggregated data into a new file.

Before any analysis is conducted explore the data. Box plots are a good way to do this.

1. Click on **Graphs** and then **Boxplot**.
2. In the **Boxplot** dialogue box click on **Simple**. The **Data in Charts** are **Summaries for Separate Variables**.
3. Click on **Define** and select **y95, y96, y97, y98, y99** as the variables that the boxes represent. Click on **OK**.

Figure 1. Boxplot of % foliage cover for 87 plots.




Notice that there are only 87 cases for each time even though there are 101 plots along all the nine lines. This is because there is some missing data, i.e., there were some surveys where there was no data collected on foliage cover.

There appears to be an increase in foliage cover in 1997 compared with earlier and later years. The other point to note about the box plots is that the data does not appear to departure much from being normally distributed. If it were we may consider a transformation. A suitable transformation for percentage, or proportion data is $y' = \arcsin\sqrt{y}$ where $y$ is the original data (note: convert a percentage to a proportion, e.g., 25% should be 0.25).
To make the graph more sensible the vertical axes for the boxplots were defined to have a minimum at 0.

There are many ways to conduct repeated measures analysis. Here is one method that is easy to conduct in SPSS.

With the new aggregated data file:

1. Click on **Analyse** and then **General Linear Model** and then **Repeated Measures**.
2. In the **Repeated Measures** dialogue box call the **Within-Subject Factor** Name **time**.
3. The Number of Levels is **5** since there are five time points.
4. Click **Add** and then **Define** and in the **repeated** dialogue box that appears move **Y95**, **Y96, Y97, Y98 and Y99** to the **Within-Subjects Variables** box.
5. Click **Options** and select **Estimate of Effect Size** and **Descriptive Statistics** as display options. In the **Factor(s)** and **Factor Interaction** box select the variable time to appear in the **Display Means For** box.
6. Click **Continue** and **OK**.

Some of the output is as follows.

**Descriptive Statistics**

| | Mean | Std. Deviation | N |
|---|---|---|---|
| Y95 | 51.7117 | 20.0059 | 87 |
| Y96 | 52.9350 | 17.9674 | 87 |
| Y97 | 59.1029 | 13.8334 | 87 |
| Y98 | 52.0605 | 13.5094 | 87 |
| Y99 | 54.9702 | 14.8634 | 87 |

**Multivariate Tests**

| Effect | | Value | F | Hypothesis df | Error df | Sig. | Eta Squared |
|---|---|---|---|---|---|---|---|
| TIME | Pillai's Trace | .687 | 45.528 | 4.000 | 83.000 | .000 | .687 |
| | Wilks' Lambda | .313 | 45.528 | 4.000 | 83.000 | .000 | .687 |
| | Hotelling's Trace | 2.194 | 45.528 | 4.000 | 83.000 | .000 | .687 |
| | Hotelling's Trace | 2.194 | 45.528 | 4.000 | 83.000 | .000 | .687 |
| | Roy's Largest Root | 2.194 | 45.528 | 4.000 | 83.000 | .000 | .687 |
| | Roy's Largest Root | 2.194 | 45.528 | 4.000 | 83.000 | .000 | .687 |

a Exact statistic

b Design: Intercept  Within Subjects Design: TIME

In this analysis repeated measures data is analysed by a multivariate test. All the multivariate tests have given the same result in this example, but generally Wilks $\Lambda$ is the most commonly used test (the Greek symbol $\Lambda$ is called Lambda). There is evidence that there is a significant time effect, Wilks $\Lambda$ = 0.313, $F_{4, 83}$ = 45.528, P < 0.001.

The next step given there is a time effect is to investigate where differences in foliage cover are occuring. One way to do this is by pairwise comparisons, where the all possible pairs in time are compared.

1. Click **Analyse**, and then **Compare Me**ans and then **Paired Samples T-Test**.
2. Click y95 and y96 as variable 1 and 2 in the **Current Selections** box.
3. Click I for y95 - y96 to appear in the **Paired Variables** box.
4. Continue selecting all pairs, i.e., **y95 - y97, y95 - y98, y95 - y99, y96 - y97** etc.
5. Click **OK**

**Paired Samples Test**

| | | Paired Difference | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference | | | | |
| | | | | | Lower | Upper | | | |
| Pair 1 | Y95 - Y96 | -1.2094 | 9.0954 | .9696 | -3.1366 | .7177 | -1.247 | 87 | .216 |
| Pair 2 | Y95 - Y97 | -7.3912 | 12.1553 | 1.3032 | -9.9819 | -4.8006 | -5.672 | 86 | .000 |
| Pair 3 | Y95 - Y98 | -.4585 | 14.0733 | 1.5002 | -3.4403 | 2.5234 | -.306 | 87 | .761 |
| Pair 4 | Y95 - Y99 | -3.3351 | 16.1485 | 1.7214 | -6.7566 | 8.644E-02 | -1.937 | 87 | .056 |
| Pair 5 | Y96 - Y97 | -6.8560 | 9.0661 | .9401 | -8.7231 | -4.9889 | -7.293 | 92 | .000 |
| Pair 6 | Y96 - Y98 | -.2207 | 11.9437 | 1.2319 | -2.6670 | 2.2256 | -.179 | 93 | .858 |
| Pair 7 | Y96 - Y99 | -3.4279 | 14.5196 | 1.4976 | -6.4018 | -.4540 | -2.289 | 93 | .024 |
| Pair 7 | Y96 - Y99 | -3.4279 | 14.5196 | 1.4976 | -6.4018 | -.4540 | -2.289 | 93 | .024 |
| Pair 8 | Y97 - Y98 | 6.3295 | 6.2830 | .6283 | 5.0828 | 7.5762 | 10.074 | 99 | .000 |
| Pair 8 | Y97 - Y98 | 6.3295 | 6.2830 | .6283 | 5.0828 | 7.5762 | 10.074 | 99 | .000 |
| Pair 9 | Y97 - Y99 | 2.9786 | 9.5105 | .9510 | 1.0915 | 4.8657 | 3.132 | 99 | .002 |
| Pair 9 | Y97 - Y99 | 2.9786 | 9.5105 | .9510 | 1.0915 | 4.8657 | 3.132 | 99 | .002 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Pair 10 | Y98 - Y99 | -3.3178 | 7.2370 | .7201 | -4.7465 | -1.8891 | -4.607 | 100 | .000 |

There are 10 pairwise comparisons and one way to control the familywise error rate is to test at the 0.05/10 level, i.e, $\alpha = 0.005$. The first obvious result is that all comparisons with 1997 are significant. This is what we would expect given the trend in the boxplot where 1997 had a higher foliage cover than in other years. The other significant comparison is 1998 and 1999. The Holm's method of testing multiple comparisons gives similar results.

Because the time intervals are equally spaced (by 1 year) polynomial contrasts can be conducted.

1. Click **Analyse**, then **General Linear Models**, then **Repeated Measures** and then **Define** as before.
2. Click **Contrast** and in the **Repeated Measures-Contrasts** dialog box change the contrast to **Polynomial**.
3. Click **Continue**
4. Click **OK**

Tests of Within-Subjects Contrasts

**Measure: MEASURE_1**

| Source | TIME | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| TIME | Linear | 276.984 | 1 | 276.984 | 1.653 | .202 |
| | Quadratic | 601.405 | 1 | 601.405 | 14.566 | .000 |
| | Cubic | 218.155 | 1 | 218.155 | 7.912 | .006 |
| | Order 4 | 2121.708 | 1 | 2121.708 | 137.683 | .000 |
| Error(TIME) | Linear | 14407.838 | 86 | 167.533 | | |
| | Quadratic | 3550.827 | 86 | 41.289 | | |
| | Cubic | 2371.400 | 86 | 27.574 | | |
| | Order 4 | 1325.265 | 86 | 15.410 | | |

There are five time periods so contrasts up to order 4 can be tested. However, the linear and quadratic contrasts are the easiest to interpret. There is no evidence of a linear trend in the foliage cover over time ($P = 0.202$) but there is evidence of a quadratic trend ($P< 0.001$). The foliage cover in the earlier years was less than in middle year (1997) where after it reduced again.

Note: Line 2 received extra possum control compared with the other lines. Therefore line 2 could be considered one "treatment" and the other lines another "treatment". The appropriate analysis is then two-way repeated measures with two within-subject factors, time and possum.

## Fisheries Bycatch of Sea Lions in SQU6T, the Squid Fishery Around the Auckland Islands

MATRIX PLOTS

## LOGISTIC REGRESSION

**Excel data file: SL1000.xls (DME CHCRO-28734)**

This example is based on a random sample of 1000 observed tows from the squid fishery in the area known as SQU 6T around the Auckland Islands. Official observers have been placed on about 15% of fishing vessels in recent years in order to record the bycatch of sea lions during fishery operations. The data here are for tows taken between June 1991 to September 1996. They were selected from a complete data set of all tows during this period, which was constructed from several component databases provided by the Ministry of Fisheries.

There are two reasons why the full data set was not used for this example. First, it is questionable whether the successive tows on one vessel provide completely independent data. This can be allowed for in an analysis, but not easily in SPSS. Second, the calculations for logistic regression are relatively slow so for the purpose of an example it seemed best to keep the sample size to 1000. Note, however, that SPSS can deal with very large sample sizes if necessary.

The data are in the Excel file SL1000.xls. The variables in order are as follows:

| | |
|---|---|
| Slion | 0 for no bycatch, 1 for the bycatch of a sea lion. |
| Vsize | the index of the size of a vessel. |
| Nation | The nationality of the vessel, coded (1) Australia, (2) China, (3) CIS, (4) Denmark, (5) Faroe, (6) Japan, (7) Korea, (8) Norway, (9) NZ, (10) Poland, (11) Russia, (12) Ukraine, and (13) USA. |
| Fyear | The fishing year coded from (1) for 1990/91 to (6) for 1995/96. |
| Season | The season of the year, coded (1) for summer (December to February), (2) for autumn (March to May), (3) for winter (June to August), and (4) for spring (September to November). |
| Target | The target fishery coded(3) for squid, (4) for scampi, and (9) for other species (Trachurus murphyi, smooth oreo and silver warehou). |
| Gear | The gear type, coded as (1) for midwater and (2) for bottom tows. |
| Tday | The time of day when tows began, coded (1) for 12 midnight to 6 am, (2) for 6 am to 12 noon, (3) for 12 noon to 6 pm, and (4) for 6 pm to 12 midnight. |
| Logdur | The natural logarithm of the average time of fishing per tow for the tows in the sample unit. |
| Logwt | The natural logarithm of the green weight of fish caught. |

Using the data, the task is to determine whether the probability of bycatch is related to one or more of the variables Vsize to Logwt.

Logistic regression assumes that the relationship between the probability of bycatch (P) and the explanatory variables takes the form

$$P = \text{Exp(Linear Combination)}/\{1 + \text{Exp(Linear Combination)}\},$$

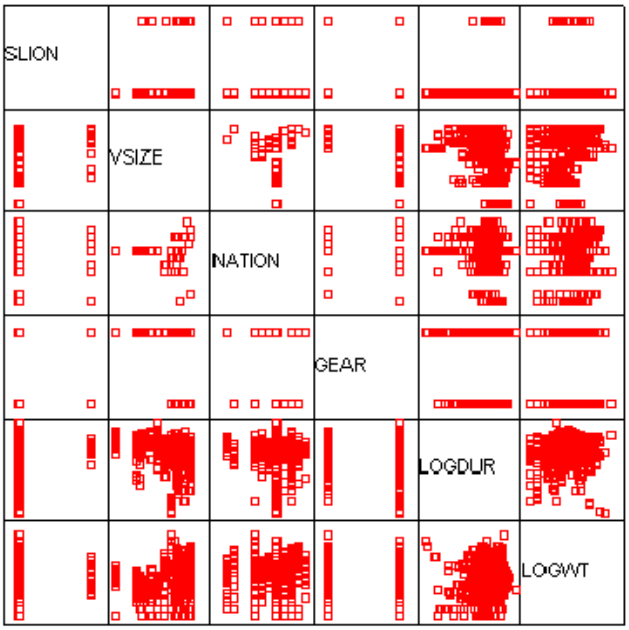
where 'Linear Combination' comprises terms that allow the mean to vary with the level of a factor and to have a linear relationship with a variable.

Here an example of a factor is Nation. The logistic regression would in this case include a term $N_i$ if a tow was on a vessel with Nationality i. This term then varies from tow to tow according to the nationality involved. An example of a variable is Vsize. For this the linear combination includes a term $\beta$ (Vsize) as with a normal multiple regression. Logistic regression is a means of estimating the effects (the $N_i$ here) for factors and the coefficients (the $\beta$ here) for variables. Various tests are then available to see whether the estimated terms are significant.

**Initial Look at Data**

Open Excel spreadsheet file SL.1000 (DME CHCRO-28734)

1. Copy and paste the data into SPSS Data editor window. Go into variable view (see bottom of screen), under name insert appropriate variable names, e.g., label the first variable **SLION**.
2. Choose **Graphs > Scatter > Matrix** and select **Slion** and the variables that describe the tow characteristics (Vsize, Nation, Gear, Logdur and Logwt). Run the procedure, which should give you the output below. Note that the more variables that you use for a matrix plot, the cruder that the plot gets. This is about as many as is reasonable.



3. Scan the matrix plot to look for any interesting patterns.
4. Return to the data window, and then back to the matrix plot option. Put in Slion and the other variables so far not considered (Fyear, Season, Target, Tday). Run the procedure again. You should obtain the plot below.
5. Again, scan the matrix plot for interesting patterns.


**Logistic Regression**

1. Return to the data window, and choose **Analyse > Regression > Binary Logistic**.
2. Choose **slion** as the dependent variable.
3. Choose **Vsize** as a covariate.
4. Choose **Nation** as a covariate. This must now be entered as a factor so choose **Categorical** and select **Nation** as being of this type. There are various ways of handling levels for categorical variables and the default is Indicator with the last category as a reference. This means that the last level of the factor is treated as the standard, from which the other levels are allowed to deviate. Choose this default, but check out the help information on the other possibilities.
5. Enter all of the other variables one by one, making them categorical if necessary.
6. For **Method**, choose **Backwards LR**, but read the help on the other options.
7. Check out the **Options**, but leave them as they are for now. Later you can go back and see what these give you.
8. Run the analysis. You should obtain the results shown below.
9. What do you conclude from this analysis?