Appendix 1: Worksheets for Practice on SPSS by Workshop Participants

WORKSHEET 1

Summarising Numerical Data in SPSS

Open the Excel spreadsheet file with the Paterson Inlet blue cod data called **PICodSize**. (DME CHCRO-28728)

1. Copy data (not labels) from columns B to H into SPSS, that is, data in columns with the headings:

Site				Summary		
Number	Year	Month	Reserve	n	Mean	SD

2. Open SPSS and paste data into the data editor window. Go into variable view (see bottom of screen), under name insert appropriate **variable** names, e.g., label the first **variable** site number.

3. Summarise the site-summary variables: n, mean and sd using **Analyse** > **Descriptive Statistics** > **Explore**. Click on the option **Plots** and select stem-and-leaf plot and a histogram as well as boxplots.

   Which plot is the most useful for displaying the data?

4. In the output window double click on the histogram for the **variable** "n" to open the chart editor. Under the option **Chart** click on axis and change the axis of the graph so that it has a maximum at 20. Add a normal curve to the graph (**Chart**, **Options** and **Display Normal Curve**).

5. Repeat the steps above but where the sites within the proposed reserve are separated from sites outside the proposed reserve. The **Factor List** is the **variable** that was called "Reserve" in the spreadsheet (1 = in the proposed reserve, 0 = outside the proposed reserve). This should produce box plots on the single graph for each **variable**.

   Is there evidence of differences in fish in sites inside and outside the proposed reserve?

   How would you test this statistically?

6. The data were collected over 4 years (1994 to 1998). Summarise the mean fish length at survey sites by year and by survey site status (i.e., whether it is inside

or outside the proposed reserve). One way to do this is to use **Analyse** > **Compare Means** > **Means**. Put the **variable** "Mean" in the **Dependent List**. The **Independent List Variables** is "Year". Click on **Next** and make "Reserve" the **variable** for the next layer.

Plot the data summarised for year and reserve status as box plots, e.g., **Graphs** > **Boxplot** > **Clustered > Data in Charts - are for Groups of Cases**.

Do the differences between mean fish length at sites within and outside the proposed reserve differ over time?

Is there anything unusual about the data in 1997? Recall there was little difference between the *average* of the fish means in 1997 for sites within and outside the proposed reserve however, the box plots show the *median* of the data.

WORKSHEET 2

SUMMARISING NUMERICAL DATA IN SPSS CONTINUED

Open the Excel spreadsheet file Dataset7 (DME CHCRO-28729), tree conditions at Ryan Creek, Heaphy Valley in 1995 and 1999. The purpose of this exercise is to compare the measures on trees for 1995 and 1999, and to look at the relationships between some of these measures.

The measures that will be considered are Folcov, Dbtop, Dbwho, Brtop and Brwho. These are given tree by tree in the file and it is these values that will be considered in this worksheet. Note however that for statistical analysis purposes the results for several trees on the same plot may not be independent.

See the definitions of the **variables** in the spreadsheet (or in the printed copy of page 1 of the data).

1. Copy and paste the values of Year, Line, Plot, Folcov, Dbtop, Dbwho, Brtop, Brwho and Stuse 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** year.

2. Summarise the results for the six **variables** Folcov to Brwho using **Analyse > Descriptive Statistics > Frequencies**. Look carefully at the various options, and make sure that you print a histogram for each of these **variables**.

3. A useful device is the Matrix Plot, which produces on one page plots of several **variables** each against each other. Use **Graphs > Scatter > Matrix** to produce one of these plots for the six **variables**. One of the options here is to use a different colour for different categories of cases. Use this option to produce matrix plots with different symbols for 1995 and 1999, and then for different lines.

4. The matrix plot with all six **variables** is a bit too cramped. Try producing a series of matrix graphs to answer the following questions:

   (a) How are Dbtop and Dbwho related, and is this the same for both years and all lines?

   (b) How are Brtop and Brwho related, and is this the same for both years and all lines?

(c) How are Dbwho, Brwho and Stuse related, are does this vary with the year or line?

5. Check that you can copy and paste graphs and text from the SPSS output window into MS Word.

WORKSHEET 3

FACTORIAL ANALYSIS OF VARIANCE

In Module 4, Section 4.5, there was a three factor analysis of variance was carried out using the **General Linear Model > univariate** option in SPSS. The dependent **variable** was the average number of birds of all species recorded in 40 5-minute counts. The factors Area (treated or control), Time (before or after poisoning), and summer (1995/96 to 1989/99) were all assumed to be fixed in this analysis. The data are in the Excel file DATSET12.xls. (DME CHCRO-28730).

For this worksheet we see how the analysis changes if the two areas are regarded as being randomly selected from a population of possible areas that could be used, and how multiple comparison tests can be used to compare factor levels.

Open the Excel spreadsheet file Dataset12 (DME CHCRO-28730)

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** area.

2. Choose the options **Analyse > General Linear Model > univariate**. Choose mcount as the dependent **variable**, and let Summer, Time and Area have fixed effects (i.e., assume that the levels of these factors in the data are all the levels of interest). Click on the **Model** button. Choose the full factorial if this is not set already.

3. Run the analysis. You should get the analysis of variance table that is part of Table 4.6 in Module 4.

4. Return to the data window and then back to the general factorial option. Change the Area factor to one with random effects, and run the analysis. Note how the analysis of variance table is now much more complicated. In particular, the error term used for F-tests now changes according to what is being tested. This demonstrates the importance of making the correct assumptions about fixed and random effects once you have more than one factor for your data.

5. Return to the data window and then the general factorial option again. Either leave Area as a random effects factor, or make it fixed again. You can now check the effects of other options in the analysis. There are many of these and

the help facility will guide you through the ones that you choose to look at. The following notes may help.

(a) With **Model** you can choose to include only some of the interactions in your model, using **Custom**. This may be appropriate if the full factorial analysis indicates that some interactions are insignificant because the estimates of the important interactions should be improved if the insignificant ones are removed. Note, however, that you should not remove an interaction while leaving a higher order interaction that includes it still in the model. For example, if the factors are A, B and C, do not remove A.B while leaving A.B.C in. Why is this?

(b) With **Contrasts** you can choose to compare the means at different levels of a factor. See the SPSS help for more information about the alternative ways of doing this. For real data you might, for example have the situation where the first level of a factor is a control and the other levels are increasing doses. You may then wish to compare levels 2, 3, 4 etc. with the first level.

(c) With **Plots** you can plot the mean levels for one factor against the levels for a second factor. Such plots indicate the type of interaction that exists, if any.

(d) With **Post Hoc** there are 18 different ways to compare the means for different factor levels, allowing for multiple testing. Browse the SPSS help facility for recommendations about which of these to use.

(e) **Covariates** can be added into the model. For example, if individual animals are the sample units, and large animals are expected to have higher levels for the response **variable**, irrespective of factor effects, then putting a measure of size in as a covariate makes an adjustment for the sizes of the animals used in the study. This is done by adding the term X into the model, where X is size. Of course, you must have a column of sizes in your data sheet to be able to do this.

(f) The **WLS** option allows the observations in the analysis to have different weights. A common reason for this is that the observations are actually mean values for samples of different sizes. In that case it is appropriate to put the sample size in as the weight **variable**.

(g) Finally, the **Options** button gives you the ability to print out mean values for different factor levels and combinations of factor levels, test for unequal variances at different factor levels, get residual plots, etc. Again, consult the SPSS help facility for more information.