

SSC-Stat 2.18 Tutorial

by Roger Stern, Sandro Leidi and Colin Grayer

Contents

1. Introduction	1
1.1. Data types	1
1.2. Sections of this guide	1
2. The data area	2
3. Simple descriptive methods	4
3.1. Analysing yields	4
3.2. To link or not to link?	6
3.3. Hidden data	7
3.4. More percentiles and proportions	8
3.5. Analysing qualitative or categorical data: factors	9
3.6. More on frequencies	11
3.7. Tidying up	11
3.8. Review of points covered so far	12
4. More descriptive statistics	13
4.1. Variate by a factor	13
4.2. Factor by a factor	15
4.3. Variate by a variate	15
4.4. Combinations of more than two variables	17
4.5. Two-way displays of the data	18
4.6. Choosing the way to display results	19
5. Organising the data	20
5.1. Unstacking	20
5.2. Stacking	21
5.3. Stacking two-way tables	21
6. The help system	22
6.1. Accessing help	22
6.2. Searching for help by keyword	25
7. In conclusion	26
7.1. Good practice guides	26
7.2. Adding a statistics package	26
7.3. Getting the right answer	28



© Statistical Services Centre
The University of Reading, UK

Getting Started

If SSC-Stat has not already been installed, run the setup file to transfer the files onto your computer's hard disk. Detailed setup instructions are provided on the CD and on the web download page at <http://www.ssc.rdg.ac.uk/software/download.html>. A 'typical' installation will include the example workbook files used in this Tutorial in the same folder, named SSC-Stat by default.

From within Excel, use **Tools** ⇒ **Add-ins**. Then use the **Browse** button to find the folder where the files were installed and select the add-in file (*SSC-Stat.xls*). You should notice that an extra item, called **SSCstat**, has been added to the Excel menu bar. If you cannot locate the SSC-Stat folder, use the Search facility on your desktop to find it. You need to go through this process only once: the add-in will automatically be available each time you subsequently load Excel.

How to use this tutorial

If you are new to Excel and computers, what should you do? The Tutorial assumes some prior knowledge of Excel. You will need to look elsewhere for introductory information on using Windows, managing files and working with Excel. Some resources are available on the CD and via the SSC website: follow the links to 'Resources' and look for 'Basic Excel Materials'.

If you already have some knowledge of Excel, you could follow the 'quick tour', which comprises Sections 1 to 3. Then learn to use SSC-Stat's help system, as described in Section 6.

The 'full tour' involves working through all the Sections and examples, ideally in sequence.

If you are experienced in both computing in general and Excel in particular, you could nevertheless gain some useful knowledge from the Tutorial. We suggest you skim through the materials in Sections 1 to 3 and start with Section 4.

This tutorial is also available in the SSC-Stat add-in via the menu selection **SSCstat** ⇒ **General** ⇒ **Tutorial (PDF)**

Typographical Conventions

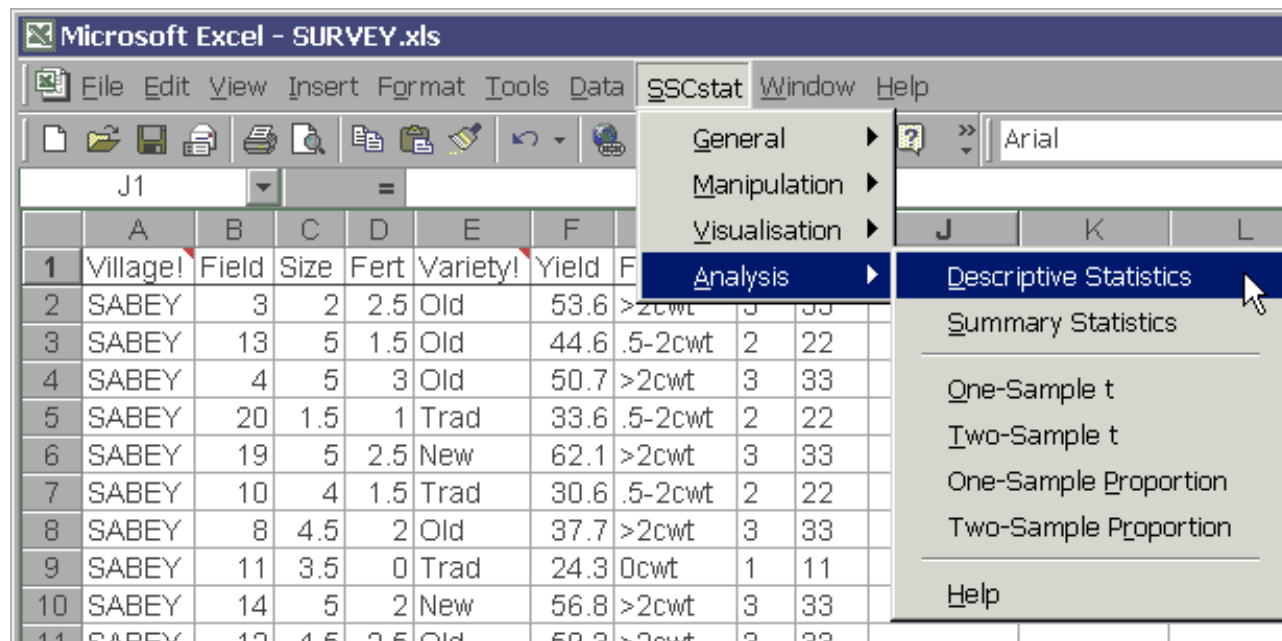
Variable names and file names are given in italics, e.g. *Yield*, *Fertiliser*, *survey.xls*.

Menu options, text on buttons and shortcut keys, and other information on dialogue boxes appear in the text in bold face, e.g. **Tools** ⇒ **Add-ins**, **[F1]**.

1. Introduction

SSC-Stat is an add-in that extends the facilities within Excel for data management and statistical analysis. It is shown in Fig. 1a.

Fig. 1a. A view of Excel showing SSC-Stat's main analysis menu



SSC-Stat is designed to support good statistical practice, which encourages exploration of the structure and relationships in the data using informative summaries, tables and graphs. Only when this has been done is it sensible to carry out more formal statistical analyses. We use SSC-Stat to support our use of Excel both to teach statistics and for data analysis.

In this tutorial we use an example dataset from a survey of rice production, stored in the *survey.xls* workbook distributed with SSC-Stat and stored in the SSC-Stat folder on your computer. The dataset is shown in Fig. 1a. We assume the reader of this tutorial is already an Excel user, so we do not need to show how to type data, nor how to use Excel for simple calculations.

1.1. Data types

The data for analysis commonly include columns of two types.

- Columns containing numbers for analysis: we call them 'variates'. These are values that investigators measure in their study. In the rice data, the columns named *Fertiliser*, *Yield* and *infected* are examples.
- Columns that indicate to which category a row of data belongs: we call these 'factors'. They may be coded as 1, 2, 3, etc, or given as text, for example Male, Female. These are values that investigators record as characteristics of each outcome: in the rice data, *Village*, *Fertlevel* and *Variety* are examples. These give structure to data.

1.2. Sections of this guide

We describe the characteristics of the survey, both to outline the data in this tutorial and to allow you to use your own examples instead.

In the survey shown in Fig. 1a there are 36 'cases' or rows of data. Nine measurements have been taken, so there are nine columns of data. The first row in Fig. 1a, gives the name of each column. The data are arranged in a single rectangular form that Excel calls a 'list'. This is the layout that we

assume in SSC-Stat. It is also a data layout that can easily be transferred to a statistics package or any other software.

We explain how to define the 'data area' as in Section 2. Then we consider analyses, using SSC-Stat's Visualisation (or graphics) menu and Analysis menu in Sections 3 and 4.

In Section 3 we look at simple descriptive analyses, when we process just single columns of the data. We first look at the yields (a continuous variable or variate), and then at the varieties used (a factor or categorical variable).

In Section 4 we examine two or more columns together. If you wish to follow the steps using your own data, you will need two categorical columns corresponding to *Village* and *Variety* in Fig. 1a. You will also need two quantitative columns like *Fertiliser* and *Yield*.

In Section 5 we illustrate the facilities in SSC-Stat for (re)organising the data. In Section 6 we describe the help facilities in SSC-Stat.

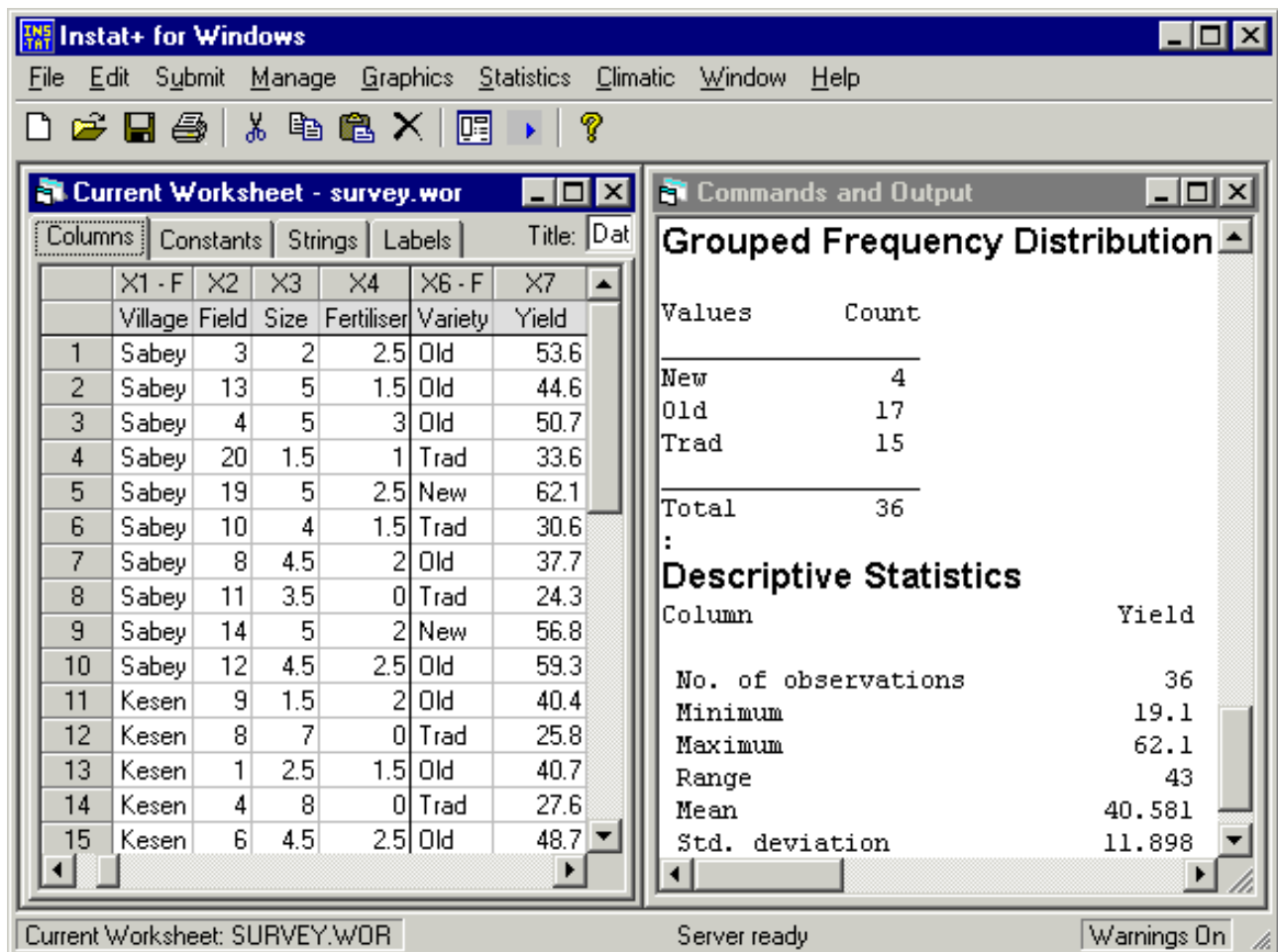
SSC-Stat is designed to complement, rather than replace the extensive facilities in Excel that support statistical analysis. They include filters, sorting of data and so on. We will use these facilities at appropriate points in this tutorial.

2. The data area

We compare one aspect of a statistics package with the use of Excel for statistical analyses, to explain why we need an option to define the data area.

In a statistics package the results appear in a window separate from that of the spreadsheet (see Fig. 2a). Hence the software knows where the data are stored all the time. The columns available for analysis are only those in the active spreadsheet.

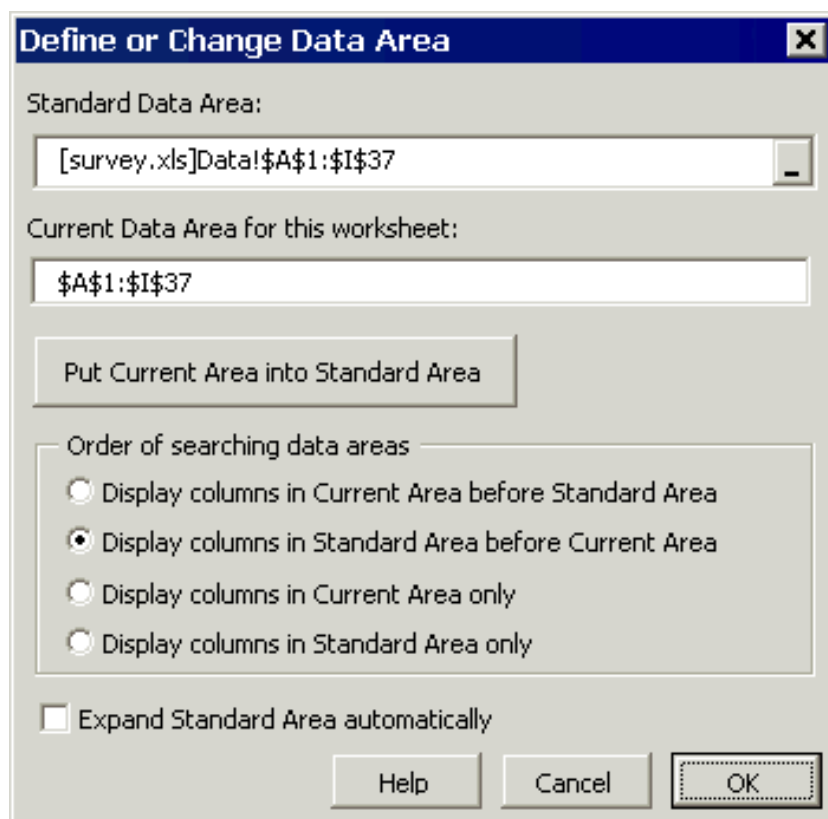
Fig. 2a. A typical statistics package (here Instat) has a separate window for results



In Excel, any results are either sent to a new sheet or simply placed alongside the data in the same sheet. Thus Excel does not differentiate clearly between data and results. To simplify the analysis it is therefore often useful to start by specifying the location of the data.

To define the data area in the survey worksheet, select (make active) any cell within the data array. Then use **General** ⇒ **Define Data Area** to bring up the SSC-Stat dialogue box (Fig. 2b). Click on the button to **Put Current Area into Standard Area** and choose the 2nd option, i.e. **Display Columns in the Standard Area before the Current Area**, as shown.

Fig. 2b. Defining the data area in SSC-Stat



This data area definition step is optional. In the next sections we will analyse the data and put results in further sheets in the workbook. If you do not define a data area, you must always return to the data sheet before doing a new analysis.

3. Simple descriptive methods

This section shows how to summarise single columns of data using the menus and dialogues.

3.1. Analysing yields

Fig. 3a. Visualisation ⇒ Boxplot

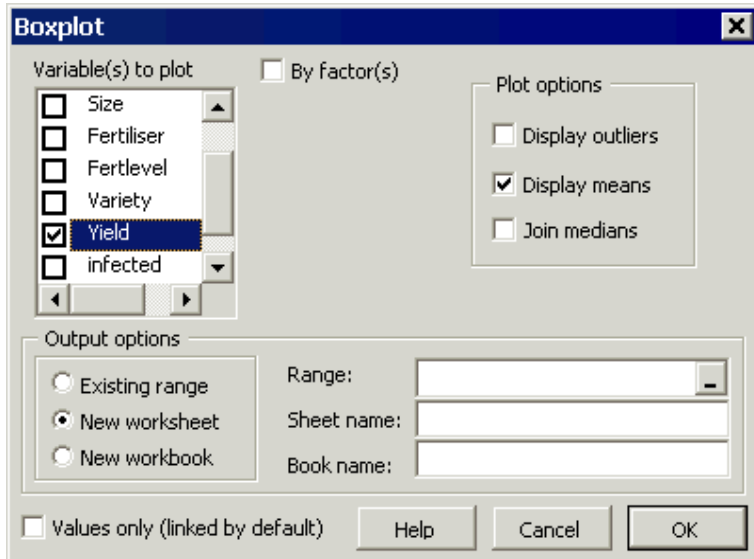
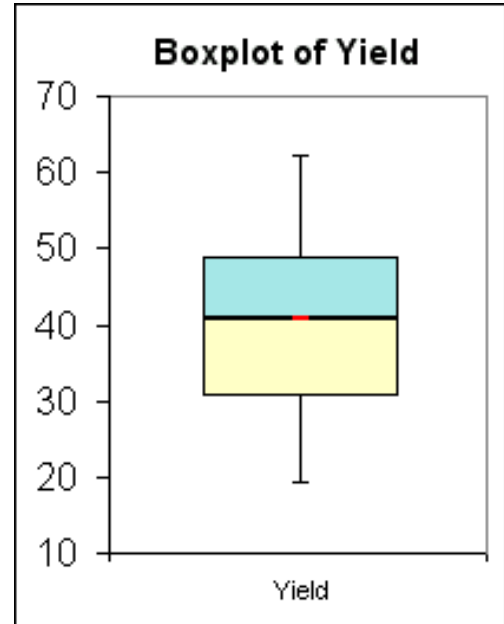


Fig. 3b. Simple boxplot



Use **Visualisation ⇒ Boxplot** as shown in Fig. 3a. Tick *Yield* as the variable to plot and press **OK**. The resulting boxplot is shown in Fig. 3b.

To help understand a boxplot, stay on this sheet and use the **Analysis ⇒ Descriptive Statistics** dialogue as shown in Fig. 3c. If there are no columns of data, i.e. if the dialogue is as in Fig. 3d, then you probably did not define the data area, as described in Section 2.¹

¹ Either define the data area, or return to the data sheet. From the data sheet you can specify that the results should be put on the same sheet as the boxplot.

Fig. 3c. Analysis ⇒ Descriptive Statistics, with Additional statistics

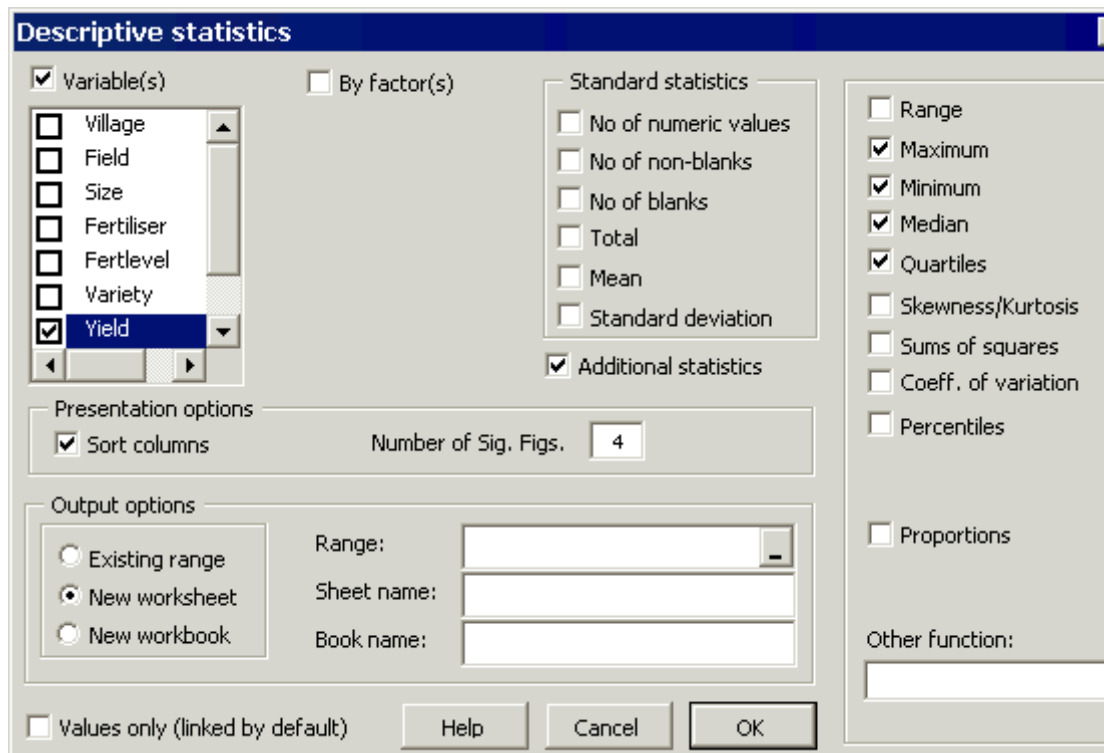


Fig. 3d. Part of the Descriptive statistics dialogue if data area is undefined



Fig. 3e. Results from the dialogue (compare with boxplot)

Descriptive Statistics	
	Yield
Minimum	19.1
Lower Quartile	30.35
Median	40.55
Upper Quartile	48.85
Maximum	62.1

Tick *Yield* as the variable to be summarised and check the box to give further statistics. Then tick the checkboxes to give the minimum, maximum, median and the quartiles, as shown in Fig. 3c. Also click the option to put results on the same sheet, so that you can compare the numerical results with the boxplot. Click **OK**.

Part of the results sheet is shown in Fig. 3e. This may be used to explain that a boxplot in Fig. 3b is a 5-number summary, where the ‘box’ shows the interquartile range. The lines – often called ‘whiskers’ – extend to the minimum and maximum.

To explore the options of this dialogue further return to the **Analysis ⇒ Descriptive Statistics** dialogue. Tick to request percentiles and type `0 25 50 75 100` in the box, as shown in Fig. 3f.

(Rest the mouse pointer on the percentiles box: a tool-tip will guide you on what to type.) Finally, change the number of significant figures to 3 and press **OK** to get the results (Fig. 3g).

Fig. 3f. Descriptive statistics again

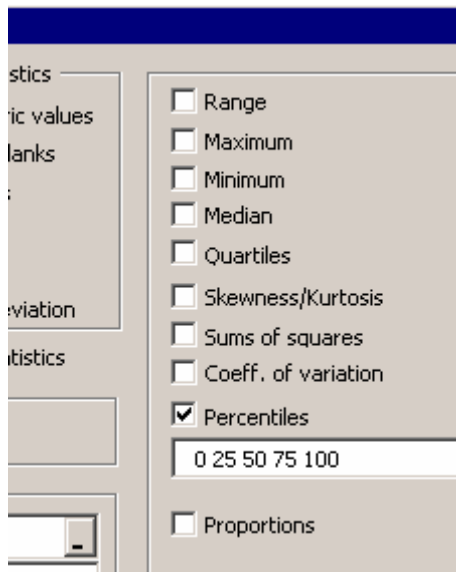


Fig. 3g. Percentiles

	K	L
Descriptive Statistics		
		Yield
0% Percentile		19.1
25% Percentile		30.4
50% Percentile		40.6
75% Percentile		48.9
100% Percentile		62.1

Check you follow the results. Note that when you ask for the 0% percentile, SSC-Stat simply gives the minimum value.

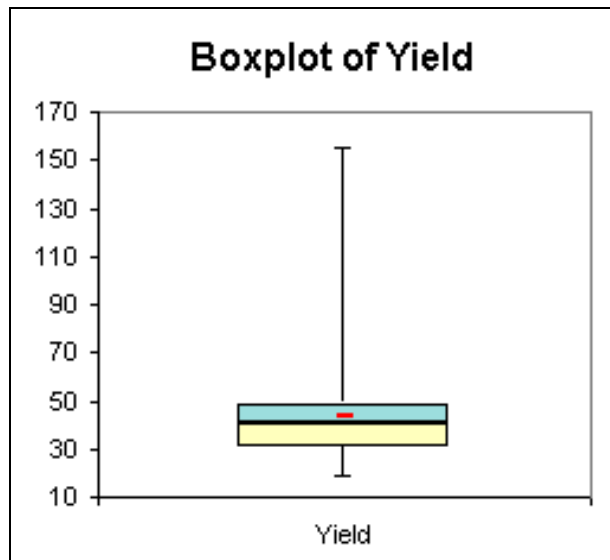
3.2. To link or not to link?

One advantage of using a spreadsheet for statistical work is that the data and results can be 'linked'. To explain what this means:

- return to the data sheet
- change the first yield from 53.6 to 153.6
- return to the results sheet.

You should see that some of the results (for example the boxplot and the maximum value) have changed automatically. The new boxplot is shown in Fig. 3h.

Fig. 3h. Boxplot with an extreme value



Now return to the data sheet and use **Visualisation** ⇒ **Boxplot** again. Tick the option to display the outliers. The dialogue should be as shown in Fig. 3i. Notice that now you do not have the option of linking. Press **OK** to give the boxplot in Fig. 3j. This shows the outlier clearly as a separate point.

Fig. 3i. Visualisation ⇒ Boxplot with outlier

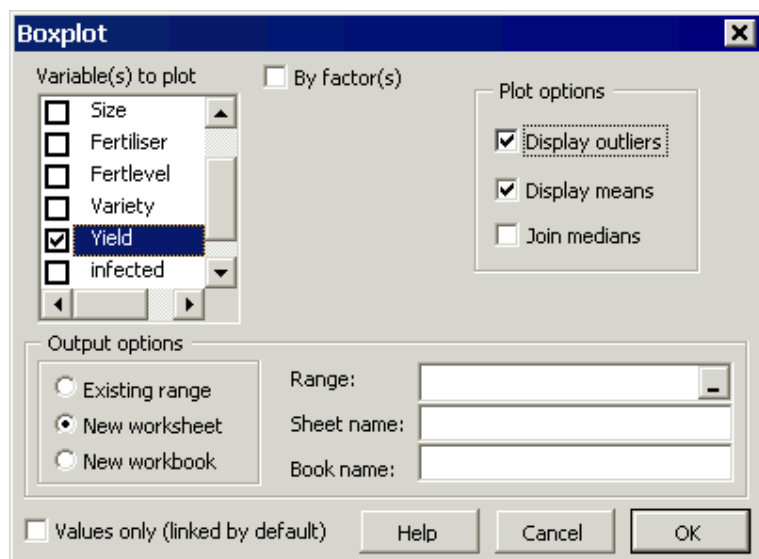
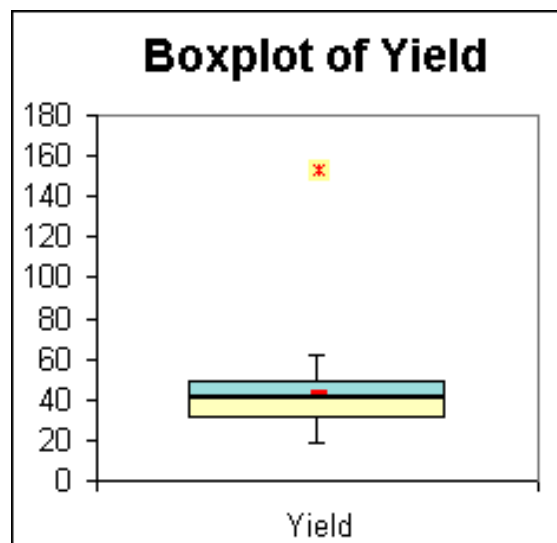


Fig. 3j. Boxplot with outlier marked



Now return to the data sheet again and correct the first value of *Yield* back to 53.6. Then look at the two result sheets. You will see that the first boxplot has changed, but the second has not.

3.3. Hidden data

In this section we use some of Excel's standard features in conjunction with SSC-Stat.

Marking a set of rows and using **Format** ⇒ **Row** ⇒ **Hide** will hide these rows from view. Columns can be hidden in a similar way. You can also hide rows in Excel's lists by using **Data** ⇒ **Filter**. The SSC-Stat dialogues always ignore hidden rows or columns. This feature adds greatly to the power of Excel for analysis. It can be used to eliminate rows where data are missing or to restrict analysis to a defined subset of the data.

If your data area also includes rows of results (which we do not recommend) then you can hide these rows before analysing the data.

Fig. 3k. Selecting one village

	A	B	C
1	Village	Field	Size
(All)		19	5
(Top 10...)		10	4
(Custom...)		8	4.5
Nanda		11	3.5
Niko		14	5
Sabey		12	4.5
11	Sabey	12	4.5
12	Kesen	9	1.5
13	Kesen	8	7
14	Kesen	1	2.5

Fig. 3l. Summary for selected village

	A	B	C
1	Descriptive Statistics		
2			
3		Yield	
4	0% Percentile	25.8	
5	25% Percentile	39.35	
6	50% Percentile	42.55	
7	75% Percentile	48.53	
8	100% Percentile	61.4	
9			
10			

As an example, we restrict the analysis to the village of Nanda. Use **Data** ⇒ **Filter** ⇒ **Autofilter**, click on the triangle in the *Village* cell (Fig. 3k) and select *Nanda*, then the only village shown. Notice that the row number changes colour to blue, indicating there are hidden rows.

Now recall the SSC-Stat **Analysis** ⇒ **Descriptive Statistics** dialogue box and click **OK**. SSC-Stat remembers the previous settings, so simply change the option to write the results to a new sheet. The results in Fig. 3l are now different because they come from the subset of 14 rows of data from Nanda village.

3.4. More percentiles and proportions

Depending on the context, percentiles other than the quartiles may be of interest. Imagine the *Yield* column contains measurement of a substance whose high levels are toxic. Thus the upper quartile is no longer a satisfactory summary because a quarter of the values are higher than it. In this context, it is common to use higher percentiles like 90% or 95%.

First, use **Data** ⇒ **Filter** ⇒ **Autofilter** again to turn off the filtering and hence use all the data.

To obtain specific percentiles, return to the **Analysis** ⇒ **Descriptive Statistics** dialogue box, tick percentiles and type in the box, then click **OK**. The relevant results are shown in Fig. 3m.

Fig. 3m. More percentiles

	A	B
1	Descriptive Statistics	
2		
3		Yield
4	90% Percentile	57.8
5	95% Percentile	59.8
6		
7		

Fig. 3n. Calculating proportions

	A	B
1	Descriptive Statistics	
2		
3		Yield
4	Proportion of values < 45	0.6667
5	Proportion of values < 55	0.8333
6		
7		

Thus only 10% of data is higher than 57.8 and only 5% of data are higher than 59.8 q/ha.²

There are also situations in which it is of interest to know the proportion of data below or above given values. Say that in order to boost productivity, the central government has established that a province qualifies for incentives if at least 60% of its yields are lower than 45 q/ha and for extra incentives if at least 90% of its yields are lower than 55 q/ha.

In the **Analysis** ⇒ **Descriptive Statistics** dialogue box select *Yield* as variable to describe, tick proportions and type in the box. Pressing **OK** gives the results in Fig. 3n.

Thus the province qualifies for incentives, as more than 60% of its yields are lower than 45 q/ha but not for the extra incentives, as less than 90% of its yields are lower than 55 q/ha.

² There is no agreed way of calculating percentiles. The differences between the different methods occur, because of the need to interpolate between successive observations to give the required percentile. These differences are slight, unless the data set is small. In the SSC-Stat help we describe the method used in Excel.

3.5. Analysing qualitative or categorical data: factors

Factors are qualitative or categorical variables, namely codes that indicate to which group a row belongs. So summaries like percentiles may be meaningless for factors, whereas frequency counts give a more relevant description.

The number of categories (or levels) of a factor and the frequency of each category are important characteristics of a dataset. How many varieties were there in our survey and how many farmers used each variety?

Fig. 3o. Analysis => Summary

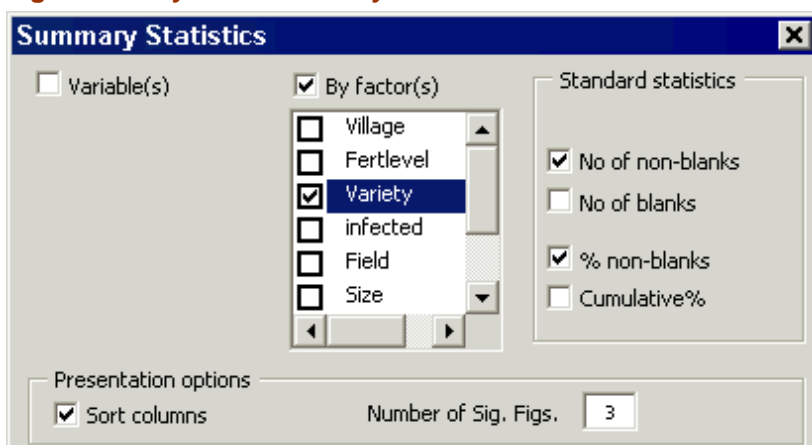


Fig. 3p. Resulting output

	A	B	C
1	Variety	CountAll	%CountAll
2	New	4	11.1
3	Old	17	47.2
4	Trad	15	41.7
5			
6			

Use the **Analysis => Summary Statistics** box as shown in Fig. 3o. Un-tick the **Variables** box, and the **By factor(s)** box will become ticked automatically. Select **Variety**, request the statistic **No of non-blanks** and **% non-blanks**. Then click **OK** to give the results in Fig. 3p.

Now use the **Visualisation => Category-Value Plot** dialogue as shown in Fig. 3q. If you click **OK** you have a simple bar chart as in Fig. 3r. Return to the dialogue and try other options, for example, a line chart.

Fig. 3q. Visualise => Category-Value Plot

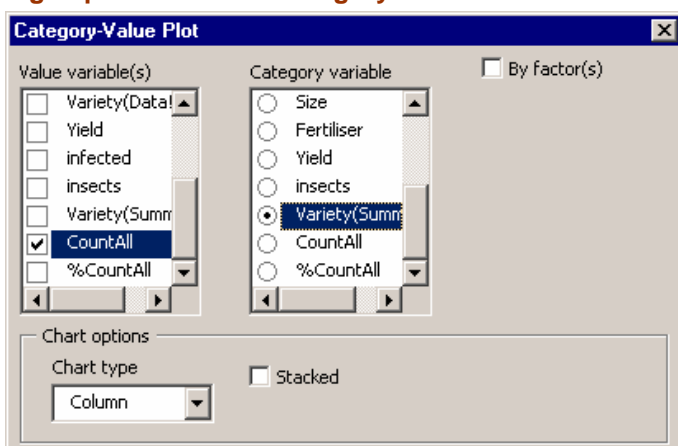
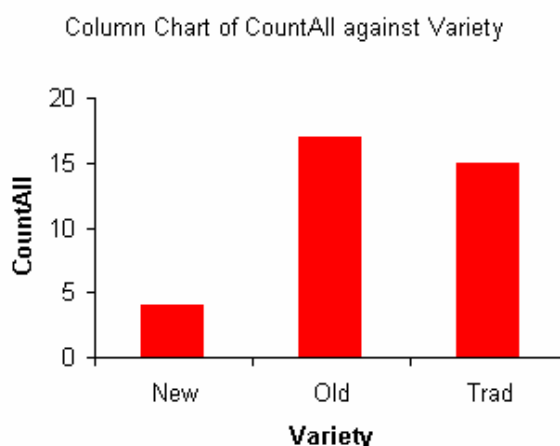


Fig. 3r. Column chart



We now make a deliberate mistake in the factor column, to see how it is reflected in the analysis:

- return to the data sheet
- change the first value in the Variety column from Old to Oldie
- use the **Analysis** ⇒ **Summary Statistics** dialogue again.

The resulting output is in Fig. 3s. It shows that any spelling mistake in the codes typed into a category, or factor column, results in a new category. The type of summary shown in Fig. 3s is a useful exploratory tool to check on this possibility, prior to the formal analysis.

Fig. 3s. Output following mistyping of a variety

	A	B	C
1	Variety	CountAll	%CountAll
2	New	4	11.1
3	Oldie	1	2.78
4	Old	16	44.4
5	Trad	15	41.7

Correct the mistake before continuing. Here, as the summary in Fig. 3s indicates there is a single wrong value it is easy to retype. In general, to do this one would use Excel's standard **Edit** ⇒ **Replace** facilities.

3.6. More on frequencies

In the survey data the column of fertiliser may be treated either as an ordinary column of data or as a factor. We look at it here as a factor. To find how many different amounts of fertiliser farmers used, return to the SSC-Stat **Analysis** ⇒ **Summary Statistics** dialogue box, tick *Fertiliser* instead of *Variety* and tick the box to give cumulative percentages as shown in Fig. 3t. The results are shown in Fig. 3u. (We used **Format** ⇒ **Cells** to specify one decimal place for the percentages.)

Fig. 3t. Analysis ⇒ Summary Statistics

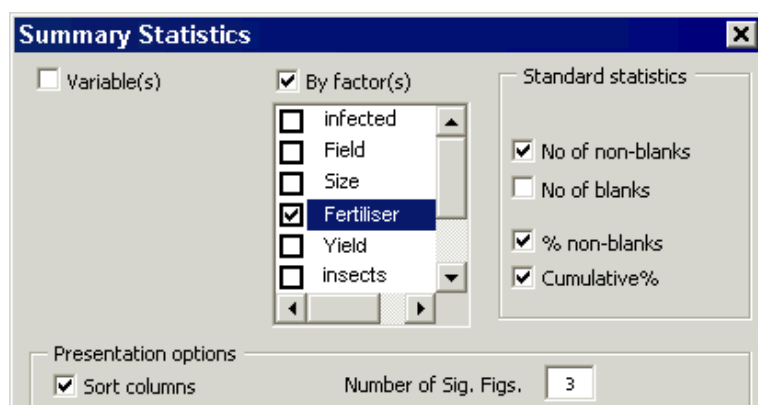


Fig. 3u. Ordered categories with percentages

	A	B	C	D
1	Fertiliser	CountAll	%CountAll	Cum%All
2	0	9	25.0	25.0
3	0.5	2	5.6	30.6
4	1	2	5.6	36.1
5	1.5	6	16.7	52.8
6	2	8	22.2	75.0
7	2.5	6	16.7	91.7
8	3	3	8.3	100.0
9				

Thus, a quarter of farmers applied no fertiliser, just over half applied no more than 150 kg/ha and three quarters applied no more than 200 kg/ha.

There are just 7 quantities of fertiliser because there are several repeats of each quantity. Though it may be tempting to consider this data column as qualitative, the repeats are the results of farmers using whole 50 kg bags. So it also makes sense to consider it as an ordinary variable, as we do in Section 4.

3.7. Tidying up

You have now produced many sheets of results. Delete the sheets before proceeding to the next section. One way is to use hold down <Ctrl> with the mouse click to mark the tab at the bottom of all the sheets except the one with the data. Then use **Edit** ⇒ **Delete** sheet.

3.8. Review of points covered so far

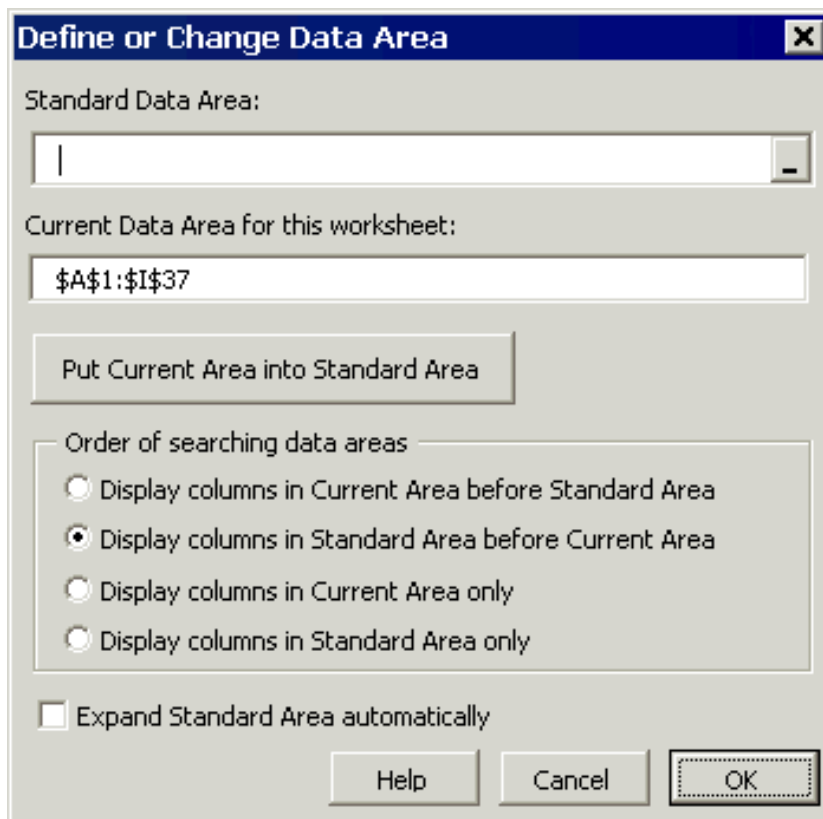
- (i) You should have found that it was easier than you expected to get the results. Hence your future analyses can concentrate on deciding *what* you should do, rather than *how* to do it.
- (ii) Why was it relatively easy to get the results? One reason is the extra facilities provided by SSC-Stat. But a second – and perhaps more important – reason is that the data were well organised before we started. They were arranged in a 'list', i.e. in columns, with each having a name at the top (as in Fig. 1a). If you want to learn more about lists, study SSC-Stat's menu **General => Help on Data Entry**.
- (iii) You have used two of the main menus in SSC-Stat, namely **Visualisation** to give graphs and **Analysis** to give summaries.
- (iv) **Visualisation => Boxplot** was used to graph the results (Section 3.1 and 3.2), and **Visualisation => Category–Value Plot** (Section 3.4) was used to graph the frequencies.
- (v) **Analysis => Descriptive Statistics** was used (Sections 3.1, 3.3 and 3.4) to give summary values of the yields, while **Analysis => Summary Statistics** was used to summarize the frequencies.
- (vi) Some other aspects commonly needed in an analysis were introduced. The feature to hide data (Section 3.3) allowed us to analyse an important subset. It can also be used to omit unwanted rows of data before the analysis. We saw that different actions are needed to summarize ordinary columns, like *Yield* (Sections 3.1 to 3.4) and category (or factor) columns like *Variety* (Sections 3.4 to 3.6).
- (vii) An analysis sometimes proceeds in stages, with the results of the first stage becoming the 'data' for the following stage. This was how we produced the column chart in Fig. 3r.

We will build on all these ideas in Section 4. In Section 5 we will introduce SSC-Stat's third main menu, when we look at how to organise data that might not be arranged as we would like.

4. More descriptive statistics

Summaries often involve more than one column and we give some examples in this section. We will be sometimes analyse the data in stages, so the dialogues are slightly clearer if you start by returning to the dialogue and clearing the data area (Fig. 4a).

Fig. 4a. General ⇒ Define Data Area to clear the standard data



4.1. Variate by a factor

From the data sheet, use the SSC-Stat **Analysis** ⇒ **Summary Statistics** dialogue, as in Fig. 4b. Tick the boxes so that both a **Variable** and a **By factor** column can be specified. Tick *Yield* as the variable to summarise, by the *Variety* factor, and request the count, mean and standard deviation.

Fig. 4b. Analysis ⇒ Summary Statistics

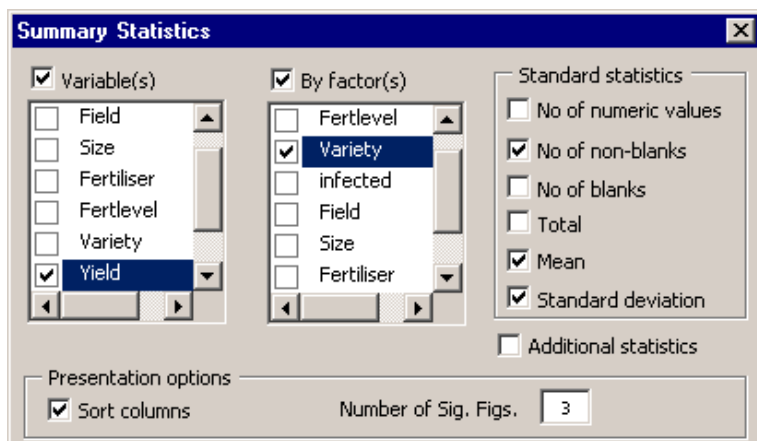


Fig. 4c. Yields from each variety

	A	B	C	D
1	Variety	CountAll	Mean	StDev
2	New	4	59.6	2.55
3	Old	17	45.4	7.13
4	Trad	15	30.0	6.52
5				
6				

From the results in Fig. 4c, it seems that the 'New' variety has higher mean yield than the other varieties and it is more stable, i.e. has smaller standard deviation. However, this is based on only four values, so we must be cautious in these conclusions.

Boxplots can also be displayed for variates according to the level of a factor. Return to the data sheet and recall the **Visualisation** ⇒ **Boxplots** dialogue box. Tick *Variety* as the **By factor**, as shown in Fig. 4d, to give results as shown in Fig. 4e.

Fig. 4d. Visualisation ⇒ Boxplot

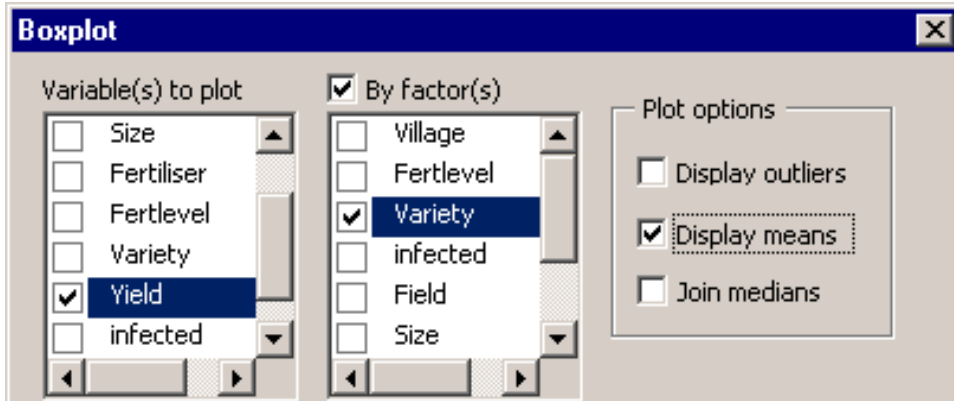
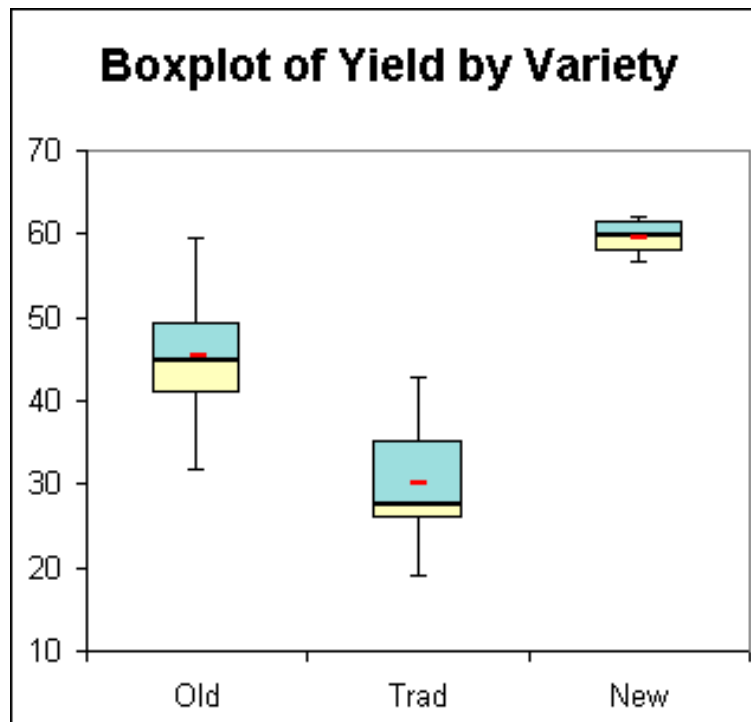


Fig. 4e. Boxplots by a factor



The range of values for the 'Trad' and 'Old' varieties is similar but the latter has higher values in general. The highest yield values are for the 'New' variety, which also has a narrower range. None of the three varieties shows an extreme value of *Yield*.

4.2. Factor by a factor

It may be interesting for the researcher to check which varieties were grown in each village, for example, to investigate the rate of adoption of the 'New' variety.

Start from the data sheet. Recall the **Analysis** ⇒ **Summary statistics** dialogue and un-tick the **Variable(s)** checkbox. Tick both *Village* and *Variety* as **By factors** (to tick two columns, hold down the <Ctrl> key when clicking on the second one) and request the number of non-blanks statistic only, as shown in Fig. 4f. The results are shown in Fig. 4g. (As an alternative, Excel's pivot tables can be used to show these results as a 2-way table. An example of a pivot table will be shown later in Fig. 4r.)

Fig. 4f. Analysis ⇒ Summary

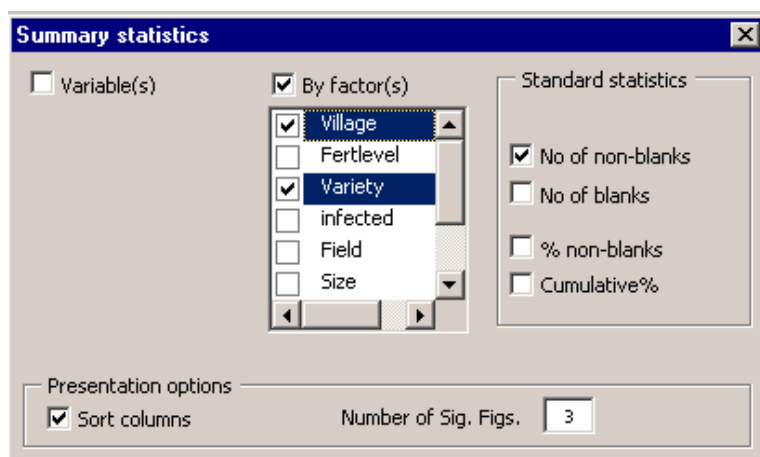


Fig. 4g. Counts of variety for each village

	A	B	C
1	Village	Variety	CountAll
2	Kesen	Old	3
3	Kesen	Trad	4
4	Nanda	New	2
5	Nanda	Old	7
6	Nanda	Trad	5
7	Niko	Old	2
8	Niko	Trad	3
9	Sabey	New	2
10	Sabey	Old	5
11	Sabey	Trad	3
12			

There are only ten rows of data (plus the header row) because the 'New' variety was not grown in the villages of Kesen and Niko.

4.3. Variate by a variate

Perhaps yield is boosted by the quantity of fertiliser applied. This could be investigated by finding the mean yield by fertiliser quantity. Return to the data sheet. In the **Analysis** ⇒ **Summary Statistics** dialogue, shown in Fig. 4h, add *Yield* again as the variable to be analysed and tick *Fertiliser* as the **By factor**. The resulting means are in Fig. 4i, and indicate that the more fertiliser is applied, the higher the yield.

Fig. 4h. Analysis ⇒ Summary Statistics

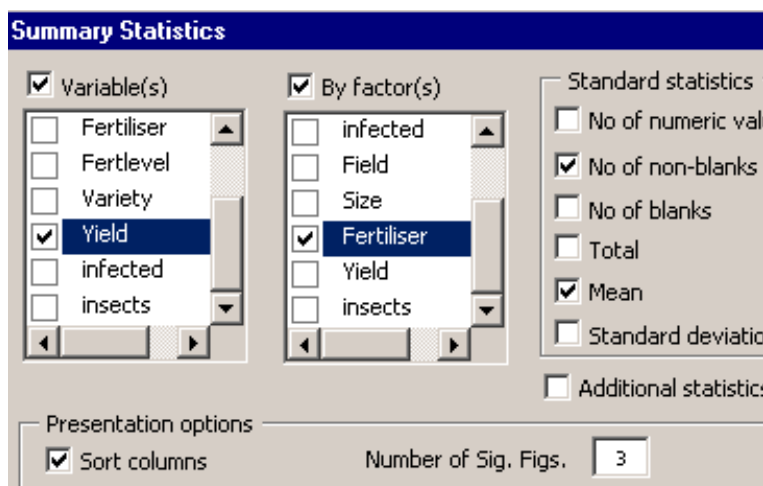


Fig. 4i. Mean for each level of fertiliser

	A	B	C
1	Fertiliser	CountAll	Mean
2	0	9	27.4
3	0.5	2	33.7
4	1	2	39.9
5	1.5	6	38.3
6	2	8	44.4
7	2.5	6	56.5
8	3	3	47.6
9			

A scatter plot might show the relationship between yield and fertiliser more clearly.

Place the cursor inside the results area in Fig. 4i, use **Visualisation** ⇒ **X-Y Scatter Plot** and complete the dialogue as shown in Fig. 4j. Choose the option to show the trend line. The graph is shown in Fig. 4k.

It seems that the mean yield increases roughly in a straight line in response to addition of fertiliser.

Fig. 4j. Visualisation ⇒ X-Y Scatter Plot

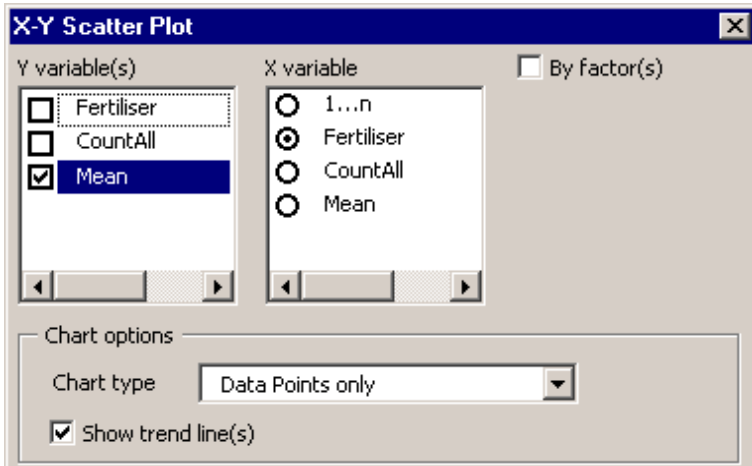
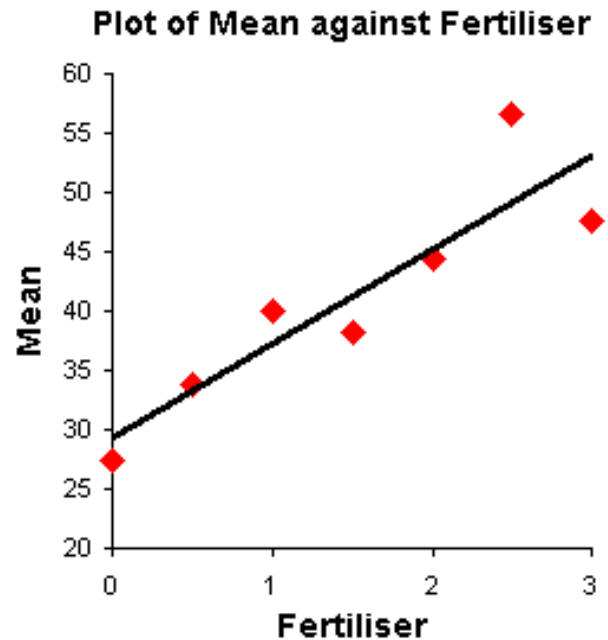


Fig. 4k. Plot using means



4.4. Combinations of more than two variables

The response of yield to fertiliser may also depend on the variety, as new varieties are bred to be more responsive to fertiliser. We first look at a graph. Use **Visualisation** ⇒ **X-Y Scatter Plot** and complete the dialogue as shown in Fig. 4l. Keep the option to show the trend line. The results are shown in Fig. 4m.

Fig. 4l. X-Y scatter plot including a factor

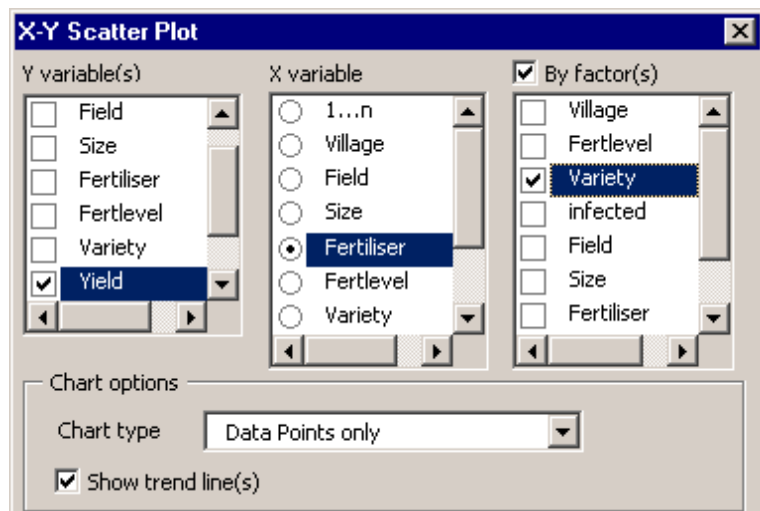
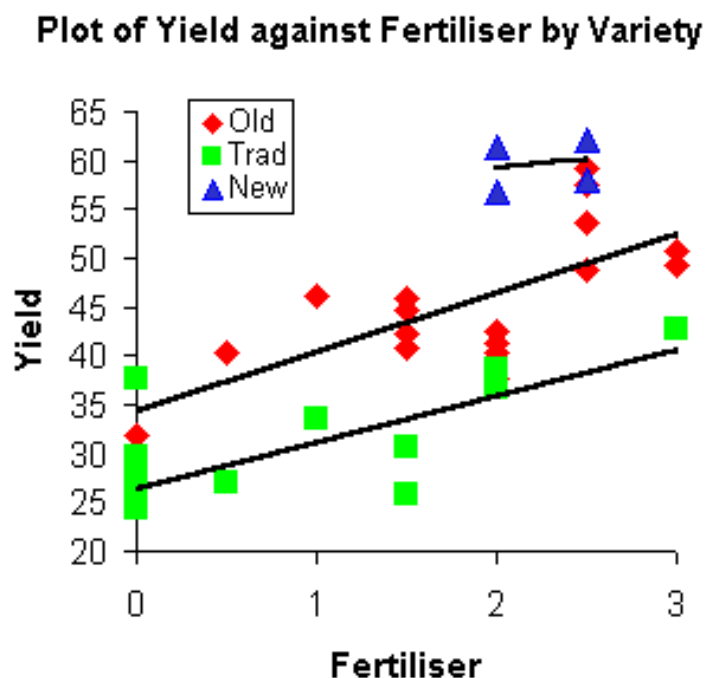


Fig. 4m. Plot with raw data and fitted lines



A numeric summary is also possible. Return to the data sheet and use the **Analysis** ⇒ **Summary Statistics** dialogue. Tick both *Variety* and *Fertiliser* as **By factor(s)**. The relevant part of the dialogue is in Fig. 4n and the results are shown in Fig. 4o.

Fig. 4n. Analysis ⇒ Summary Statistics

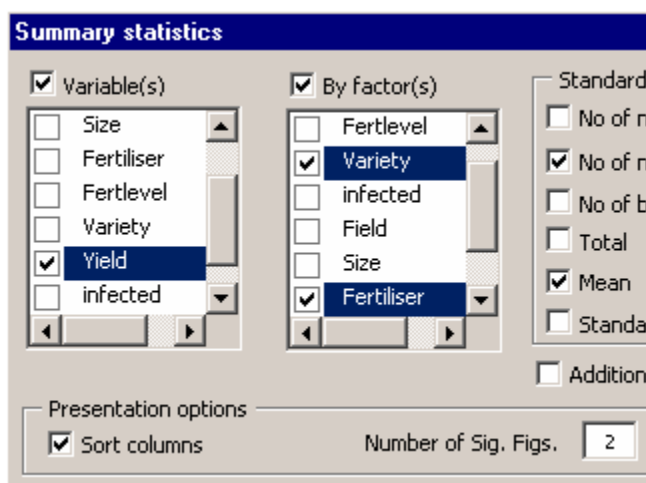


Fig. 4o. Results for Analysis ⇒ Summary

	A	B	C	D
1	Variety	Fertiliser	CountAll	Mean
2	New	2	2	59
3	New	2.5	2	60
4	Old	0	1	32
5	Old	0.5	1	40
6	Old	1	1	46
7	Old	1.5	4	43
8	Old	2	4	40
9	Old	2.5	4	55
10	Old	3	2	50
11	Trad	0	8	27
12	Trad	0.5	1	27
13	Trad	1	1	34
14	Trad	1.5	2	28
15	Trad	2	2	38
16	Trad	3	1	43
17				

While it is clear that the 'New' variety is grown with at least 200 kg of fertiliser, it is less easy to see in this table how the other two varieties respond to fertiliser application. We now look at a graph to display these summary values.

Position the cursor inside the results list shown in Fig. 4o and use the **Visualisation** ⇒ **X-Y Scatter Plot** dialogue box; select *Mean* as the **Y Variable**, *Fertiliser* as the **X Variable** and *Variety* as the **By factor**, as shown in Fig. 4p. Keep the option to draw lines. The results, after some customisation, are shown in Fig. 4q.

Fig. 4p. Plot using means

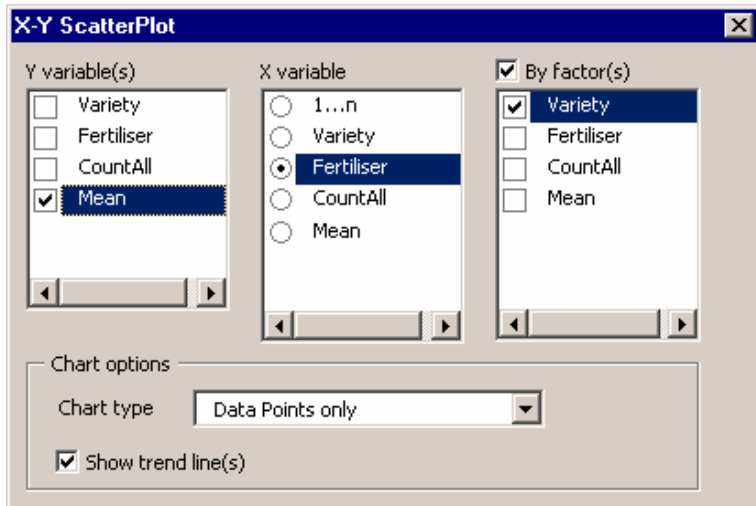
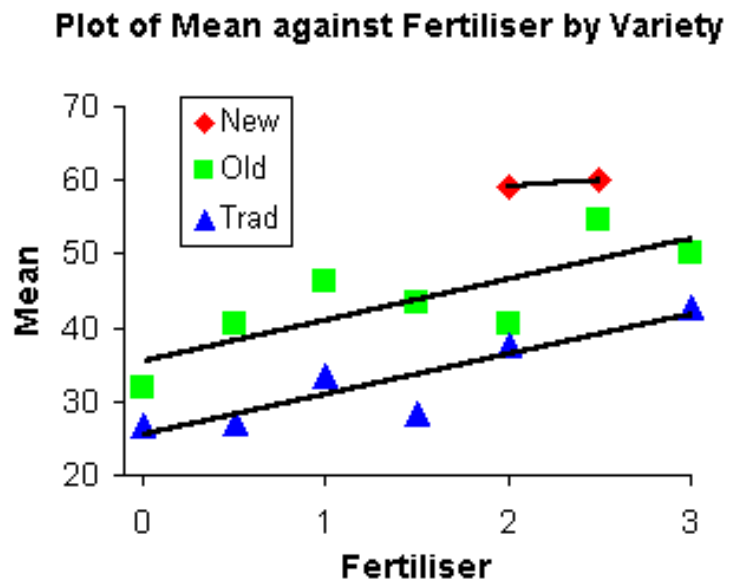


Fig. 4q. Plot to compare with Fig. 4m



The lines are similar to those given in Fig. 4h where the graph showed all the data points rather than the mean. Are the trend lines the same? If not, then which ones should be reported?

4.5. Two-way displays of the data

The summary in Fig. 4o made it easy to use to produce the graph in Fig. 4q, but is not easy to use as it stands. Excel's **Data** ⇒ **Pivot Table** and **Pivot Chart wizard** is a very powerful and useful facility for statistical work. It is described in standard guides on Excel. It is not part of SSC-Stat and is therefore not covered in this tutorial. If Excel's pivot tables are new to you then investigate them further. Fig. 4r shows the same summary as shown in Fig. 4o, but in tabular form.

Fig. 4r. Result from using Excel's Data ⇒ Pivot Table menu

	A	B	C	D	E	F	G	H	I
1									
2									
3	Average of Yield	Fertiliser							
4	Variety	0	0.5	1	1.5	2	2.5	3	Mean
5	New					59.1	60.1		59.6
6	Old	31.8	40.4	46.2	43.3	40.5	54.8	50.0	45.4
7	Trad	26.9	27.0	33.6	28.2	37.7		42.7	30.0
8	Mean	27.4	33.7	39.9	38.3	44.4	56.5	47.6	40.6

SSC-Stat does include a two-way display of either the data or the results. If, you select a cell in the list in Fig. 4o, and use **Manipulation** ⇒ **Unstack (Two-Way)**, SSC-Stat will re-arrange the results in a two-way display roughly as shown in Fig. 4r.

While Excel's pivot tables provide a two-way display of the summary values, you can use SSC-Stat to look at the raw data in a two-way array with the dialogue box shown in Fig. 4s. The data are

displayed as shown in Fig. 4t. For example with zero fertiliser we can see that the mean yield of 31.8 q/ha for the 'Old' variety is based on a single number, while the mean of 26.9 for the 'Trad' variety is the mean of 8 numbers whose individual yields varied from 19.1 to 37.6 q/ha.

Fig. 4s. Manipulation ⇒ Unstack (Two-Way)

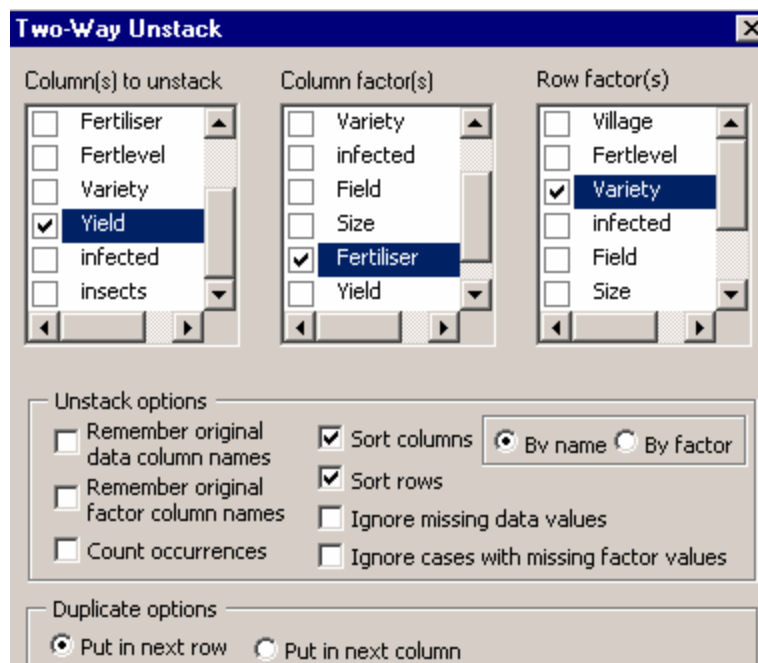


Fig. 4t. A two-way display of the raw data

	A	B	C	D	E	F	G	H
1	Variety	Fert=0	Fert=0.5	Fert=1	Fert=1.5	Fert=2	Fert=2.5	Fert=3
2	New					56.8	62.1	
3	New					61.4	58.1	
4	Old	31.8	40.4	46.2	44.6	37.7	53.6	50.7
5	Old				40.7	40.4	59.3	49.3
6	Old				42.2	41.3	48.7	
7	Old				45.8	42.4	57.4	
8	Trad	24.3	27	33.6	30.6	36.6		42.7
9	Trad	25.8			25.8	38.7		
10	Trad	27.6						
11	Trad	19.1						
12	Trad	26.3						
13	Trad	24.7						
14	Trad	29.6						
15	Trad	37.6						

4.6. Choosing the way to display results

The same tasks can be performed in both the **Analysis ⇒ Describe** and the **Analysis ⇒ Summary** dialogues, so why does SSC-Stat have two different dialogues? The difference is in how the results are displayed.

The **Describe** dialogue displays results in a tabular format, which is convenient for direct inclusion into a report without further manipulation, as you have seen in figures in Section 3.1 (e.g. Figs 3e and 3g). The results from the **Describe** dialogue can be considered as the final stage of descriptive analysis.

Instead, the **Summary** dialogue displays results in a list format. Though not normally included in reports, a list format is needed if the results themselves are to be re-used for, say, for displaying

charts as we showed in Sections 3.5, 4.3 and 4.4. You can think of results from the **Summary** dialogue as an intermediate stage of descriptive analysis.

5. Organising the data

Data are often entered in Excel spreadsheets in the format most convenient to the data clerk. The format may mimic the field form thus making the data entry process easier and quicker. Or different operators may have entered data in separate sheets.

Imagine the yields had been entered separately for each village. Once copied onto the same sheet they are likely to have the layout we are going to produce in the next section. Then they would need to be manipulated to be in 'list' format.

5.1. Unstacking

Return to the datasheet and use SSC-Stat's **Manipulation** ⇒ **Unstack** dialogue. Unstack the *Yield* column by the factor *Village* as shown in Fig. 5a. The result is shown in Fig. 5b.

This tabular format is not suitable for efficient statistical work, as SSC-Stat works on lists and so do many of Excel's own statistical features, described in Section 1.2. So, if the data start as in Fig. 5b, then we need to 'stack' *Yield* by *Village*.

Fig. 5a. Manipulation ⇒ Unstack

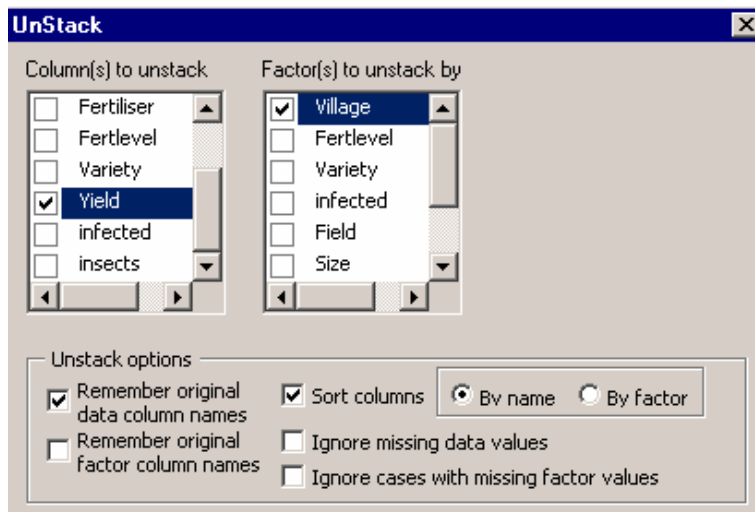


Fig. 5b. Yields in separate columns

	A	B	C	D
1	Yield:Kesen	Yield:Nanda	Yield:Niko	Yield:Sabey
2	40.4	36.6	26.3	53.6
3	25.8	57.4	24.7	44.6
4	40.7	42.7	40.4	50.7
5	27.6	49.3	31.8	33.6
6	48.7	46.2	29.6	62.1
7	27	42.2		30.6
8	19.1	41.3		37.7
9		37.6		24.3
10		58.1		56.8
11		45.8		59.3
12		38.7		
13		42.4		
14		25.8		
15		61.4		
16				

5.2. Stacking

Stay in the result sheet of Fig. 5b, and use SSC-Stat's **Manipulation** ⇒ **Stack** dialogue. Select all 4 columns to stack, change the default name of the data column to *Yield* and of the factor column to *Village* as shown in Fig. 5c. Part of the results are shown in Fig. 5d.

Fig. 5c. Manipulation ⇒ Stack

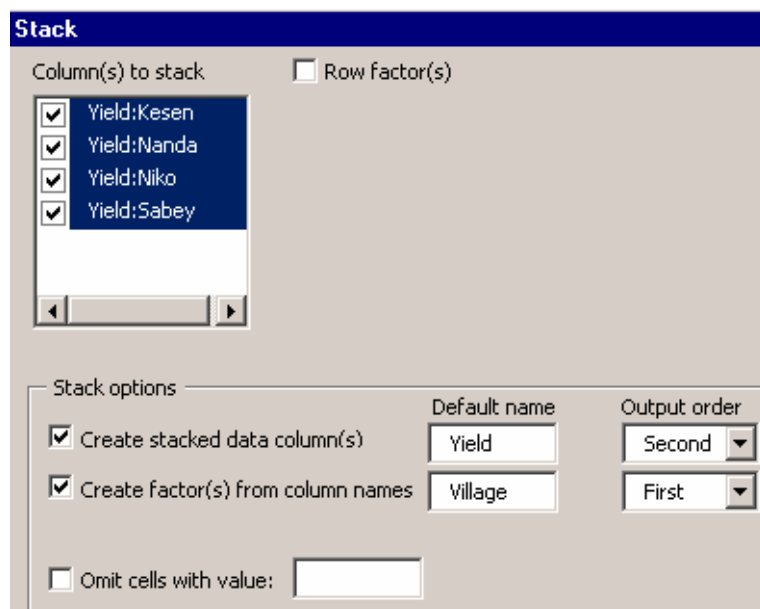


Fig. 5d. Data stacked in 'list' format

	A	B	
1	Village	Yield	
2	Kesen	40.4	
3	Kesen	25.8	
4	Kesen	40.7	
5	Kesen	27.6	
6	Kesen	48.7	
7	Kesen	27	
8	Kesen	19.1	
9	Nanda	36.6	
10	Nanda	57.4	
11	Nanda	42.7	
12	Nanda	49.3	
13	Nanda	46.2	
14	Nanda	42.2	
15	Nanda	41.3	

5.3. Stacking two-way tables


Sometimes we also have to stack a two-way tabulation, as in Fig. 4t. Place the cursor in the result sheet with the data in Fig.4t, then recall the **Manipulation** ⇒ **Stack** dialogue. Tick all columns to stack except for *Variety*. You can select a block of adjacent columns in the dialogue by holding down the <Shift> key . Click the row factor option and select *Variety*. In the right-hand side of the dialogue (Fig. 5e), notice how you can change the order of output for the stacked columns. The third option on the left-hand side of the stack option frame is for including trailing blanks. There are 6 blank cells in this two-way table, which by default are not included when stacking. Should you need to keep blanks, tick this option. The results are in Fig. 5f.

Fig. 5e. Stack including row factor

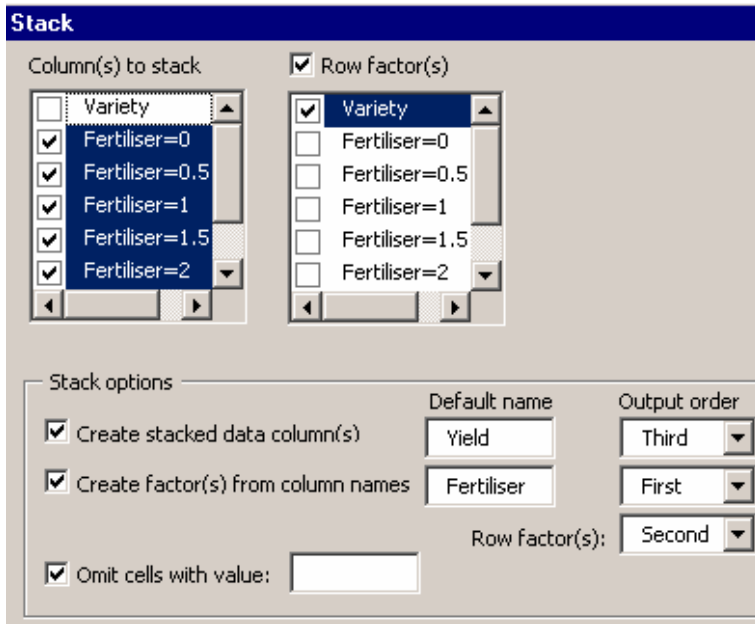


Fig. 5f. Data back in 'list' format

	A	B	C
1	Fertiliser	Variety	Yield
2	0	Old	31.8
3	0	Trad	24.3
4	0	Trad	25.8
5	0	Trad	27.6
6	0	Trad	19.1
7	0	Trad	26.3
8	0	Trad	24.7
9	0	Trad	29.6
10	0	Trad	37.6
11	0.5	Old	40.4
12	0.5	Trad	27
13	1	Old	46.2
14	1	Trad	33.6
15	1.5	Old	44.6

6. The help system

The process of analysing data involves four steps: data entry, data management, visualisation and analysis. For each step, we give general recommendations in the help system for using Excel and the facilities provided by SSC-Stat.

6.1. Accessing help

Fig. 6a. Help is available in each submenu

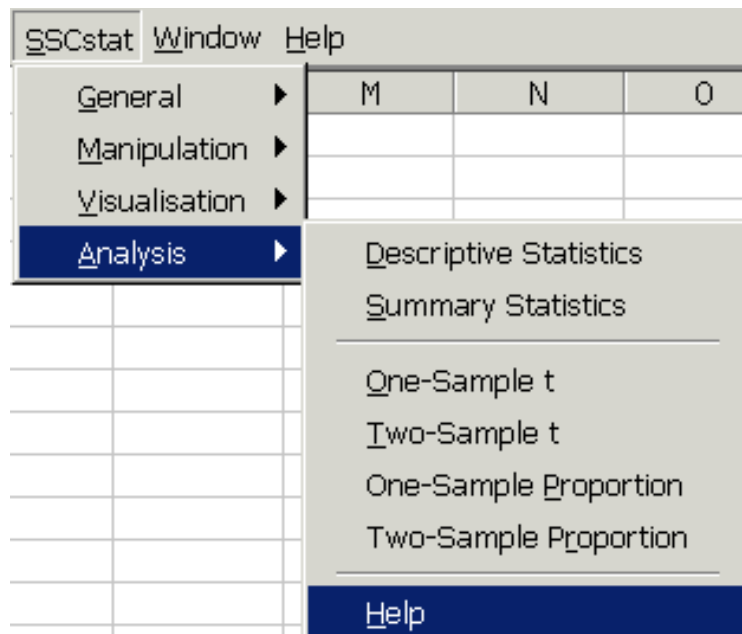
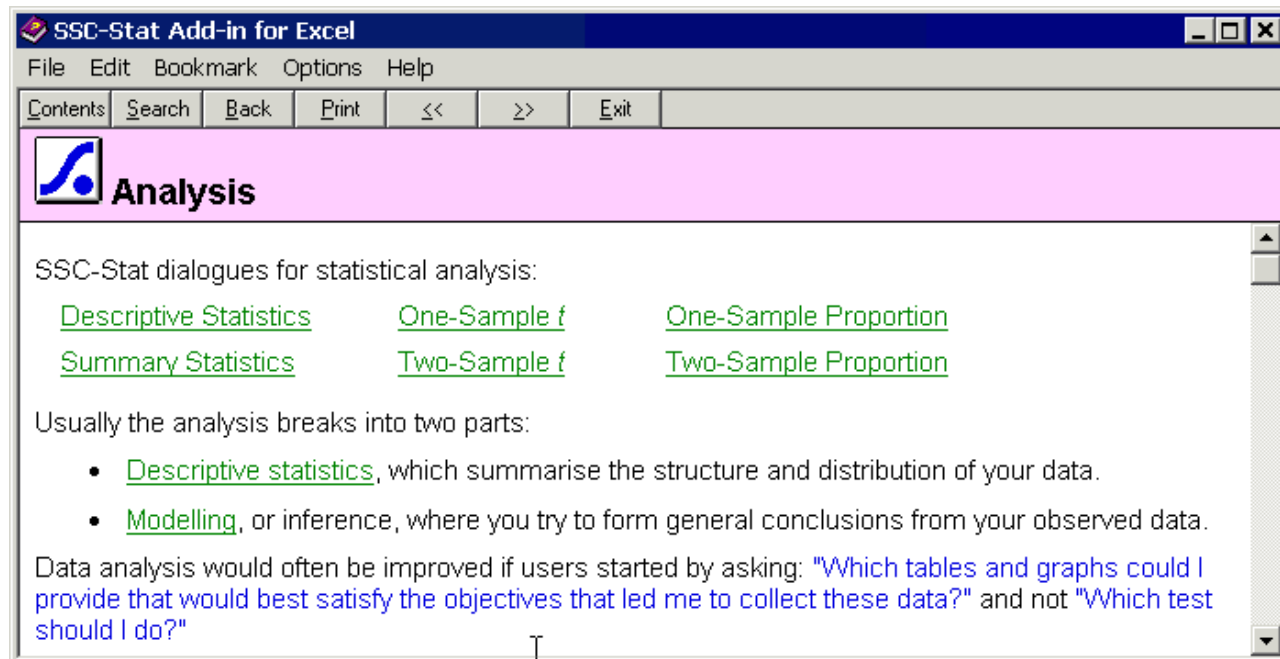


Fig. 6b. Help on Analysis



The help system appears in its own window. You can access it in three ways.

1. By pressing the **Help** button at the bottom of each dialogue box. This opens the section of on-line help relevant to the task in question. That is, pressing the **Help** button while in the **Boxplot** dialogue box shows the help for boxplots. So if you already know which task you need, access the help this way.
2. By selecting the appropriate **Help** submenu item in the SSC-Stat menu, as shown in Fig. 6a. This opens the sections of on-line help relevant to all the topics contained in the submenu. For example, the help accessed in Fig. 6a will show 6 topics corresponding to the 6 tasks available from the **Analysis** submenu (Fig. 6b). So if you know roughly what you want to do but would like help in choosing a task, access help this way.
3. By selecting the option **Overview of SSC-Stat** from the **General** submenu, as shown in Fig. 6c. This opens the help system at its top level, as in Fig. 6d, which shows all subsections that are accessible separately in the other submenus. So if you would like a tour of SSC-Stat features and pointers to the nature of statistical work, access the help this way.

Fig. 6c. General help

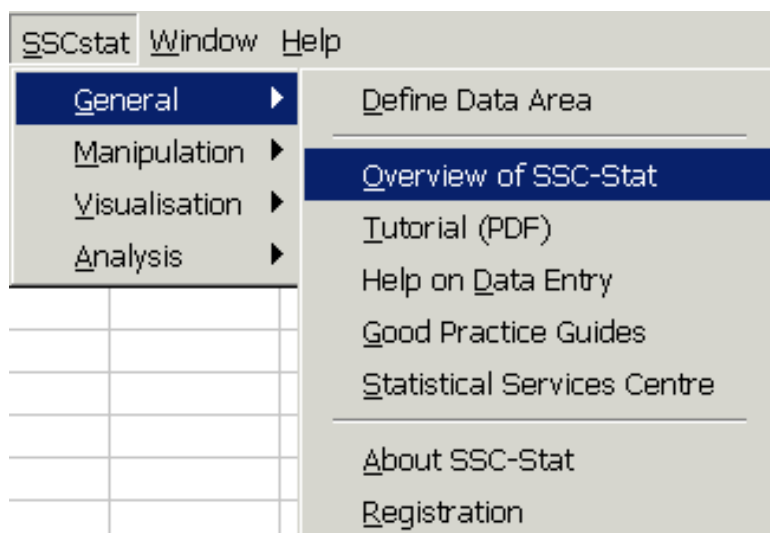
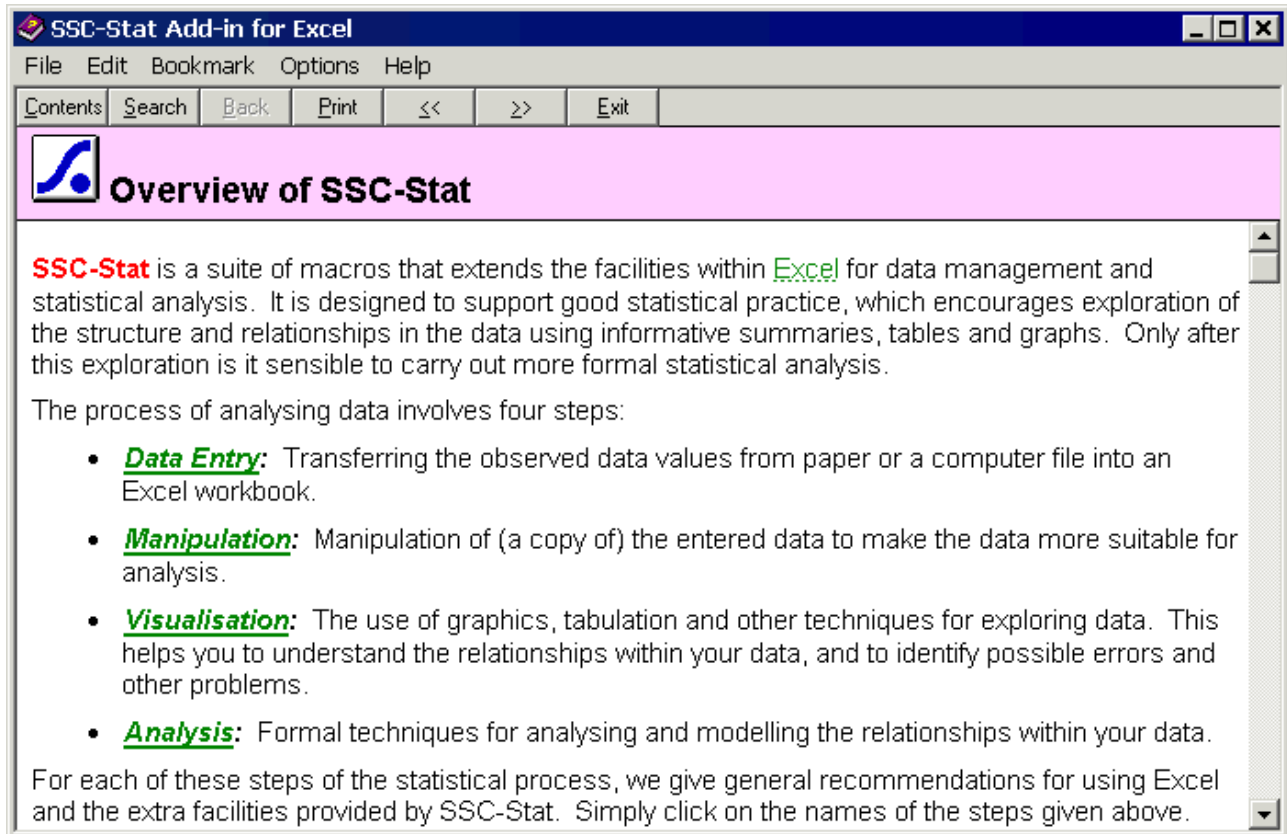
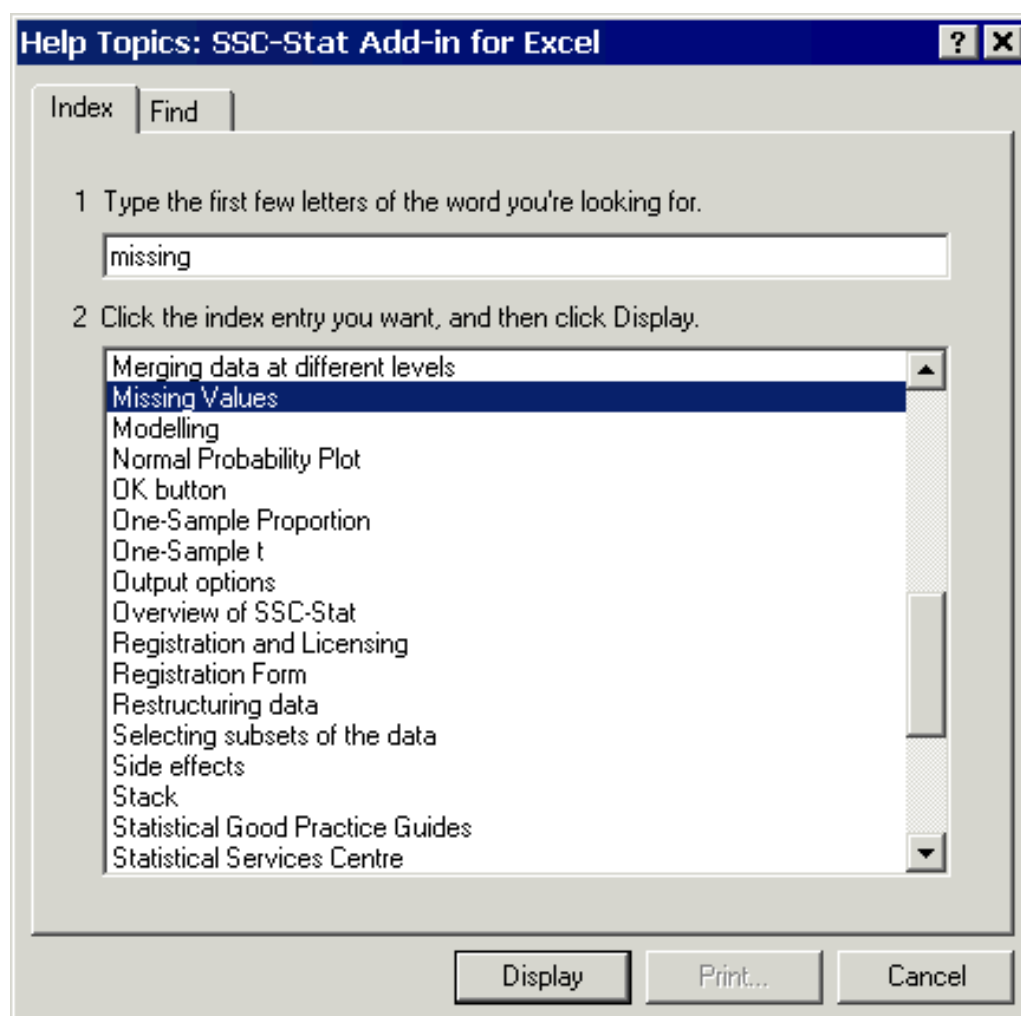


Fig. 6d. Part of the resulting help screen



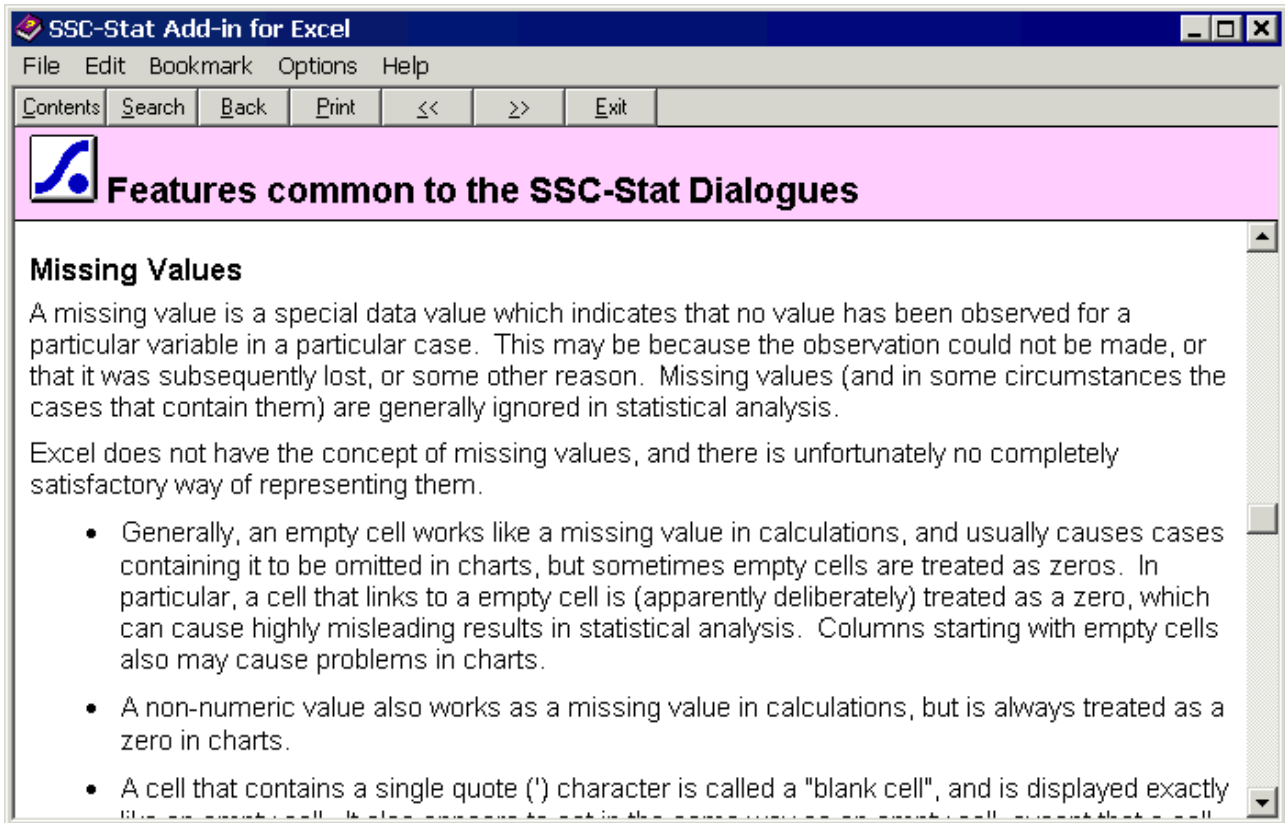
6.2. Searching for help by keyword

Fig. 6e. Searching for a help topic



Sometimes you need help about topics that are not specifically mentioned in a heading of the help files. In this case you can search help by using keywords (see Fig. 6e).

Say you want information about how to represent with missing values in Excel. From within any help window press the **Search** button to activate the **Help topics** dialogue and while on the **Index** tab, type 'missing' in the top box as shown in Fig. 6e. You will see that index entry for **Missing Values** is highlighted. Press the **Display** button to open the relevant section of the Help file (shown in Fig. 6f).

Fig. 6f. Help on missing values

Our aim in this help is to support you in becoming more productive in your statistical work. Users may feel – often correctly – that there is more useful information in their data than they have revealed by their analysis. We want to assist those who wish to exploit their data fully. We would welcome your views about SSC-Stat. Please contact us via e-mail at <statistics@lists.rdg.ac.uk>.

7. In conclusion

This section is intended more for reading than practice, though you are welcome to try some of the tasks we describe. We cover a range of issues that affect those who use Excel for their statistical work.

7.1. Good practice guides

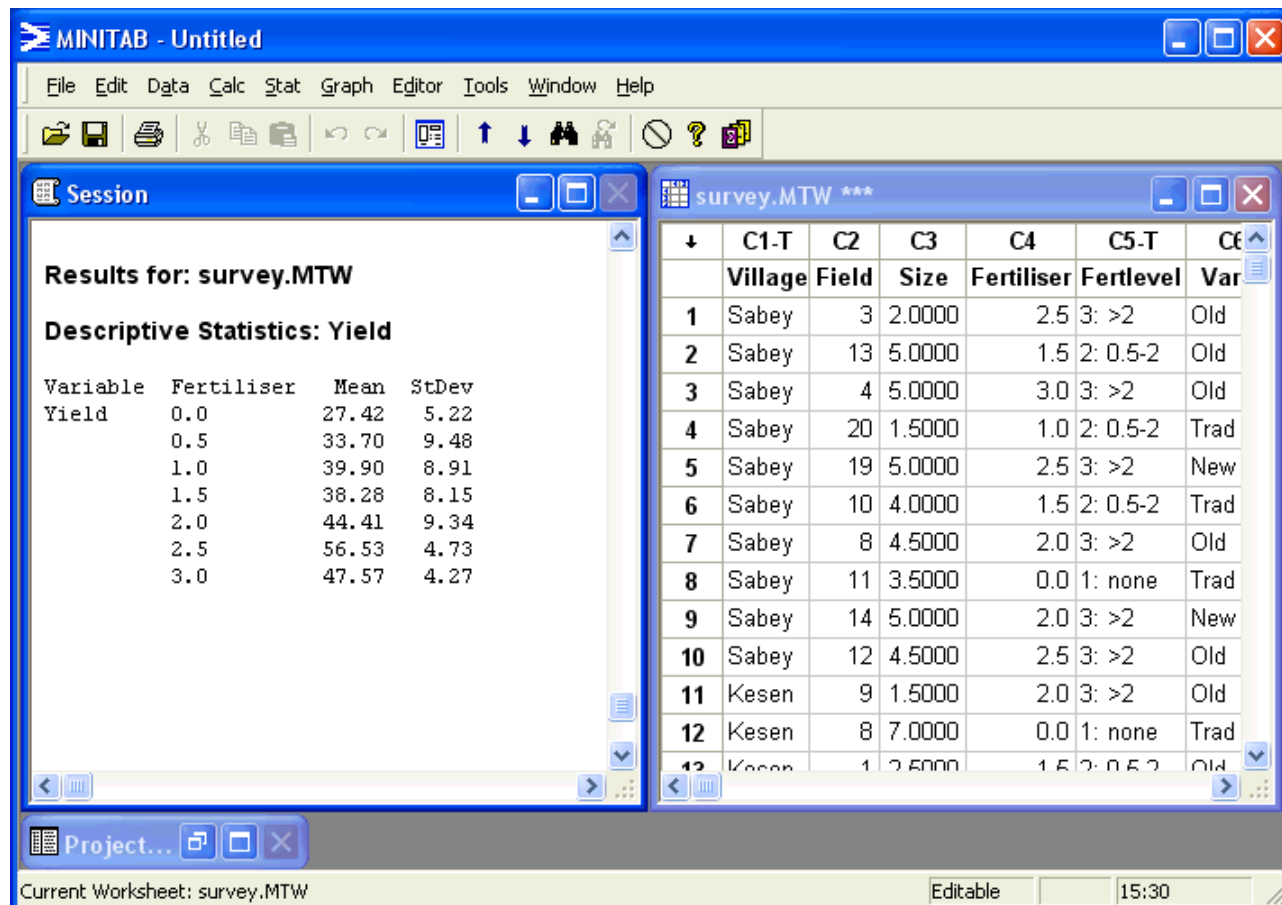
The Statistical Services Centre provides a set of statistical good-practice guidelines in printable, booklet format and as online help files. Some of the guidelines are fairly specific to Excel, such as 'Excel for Statistics: Tips and Warnings' and 'Disciplined Use of Spreadsheet Packages for Data Entry'. Others are broader in scope, such as 'Informative Presentation of Tables, Graphs and Statistics' and 'Confidence and Significance: Key Concepts of Inferential Statistics'.

The full list is available online within SSC-Stat from the menu **General** ⇒ **Good Practice Guides**. Printable versions are available on the CD and from the SSC web site: <http://www.ssc.rdg.ac.uk/>

7.2. Adding a statistics package

If you need more statistical facilities than are offered by Excel, it is easy to add a statistics package. In many packages the menus you will find are similar to those in SSC-Stat. We have called them Manipulation, Visualisation and Analysis. In Fig. 7a we show the menus of a popular package called Minitab. There are the standard Windows menus, plus the corresponding menus to organise the data, to draw graphs and to do statistical analyses. Minitab calls these menus Manip, Graph and Stat respectively.

Fig. 7a. Statistics packages have similar menus



Most statistics packages also have dialogues similar to those in SSC-Stat, which makes the transfer even easier. For example, in Fig. 7b we show the dialogue in SSC-Stat for one-sample *t*-tests. The corresponding Minitab dialogue is shown in Fig. 7c.

Fig. 7b. SSC-Stat dialogue for one-sample *t*-tests

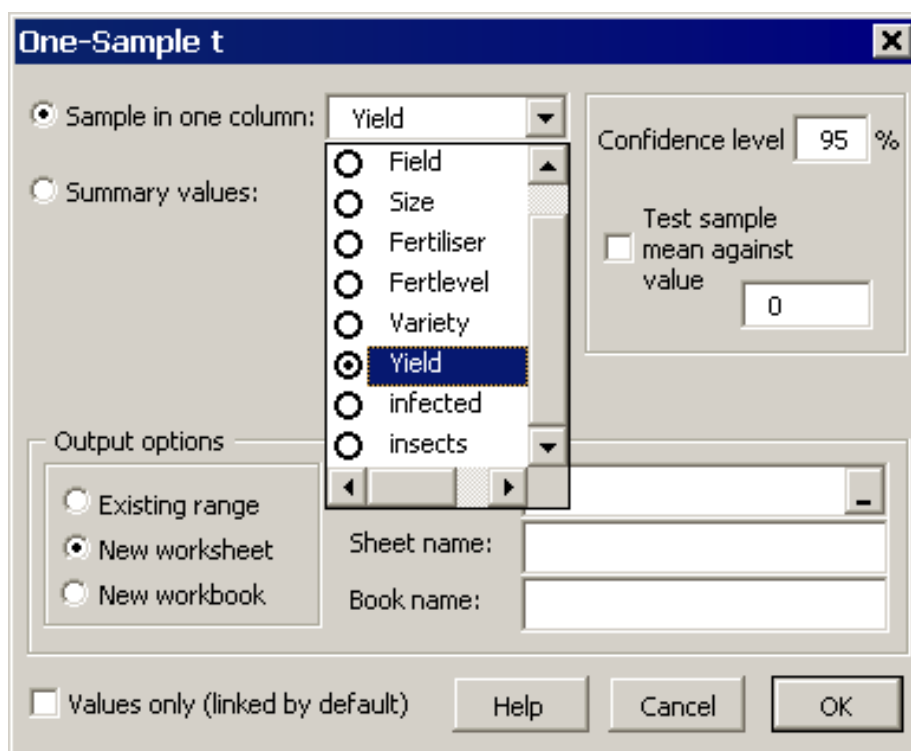
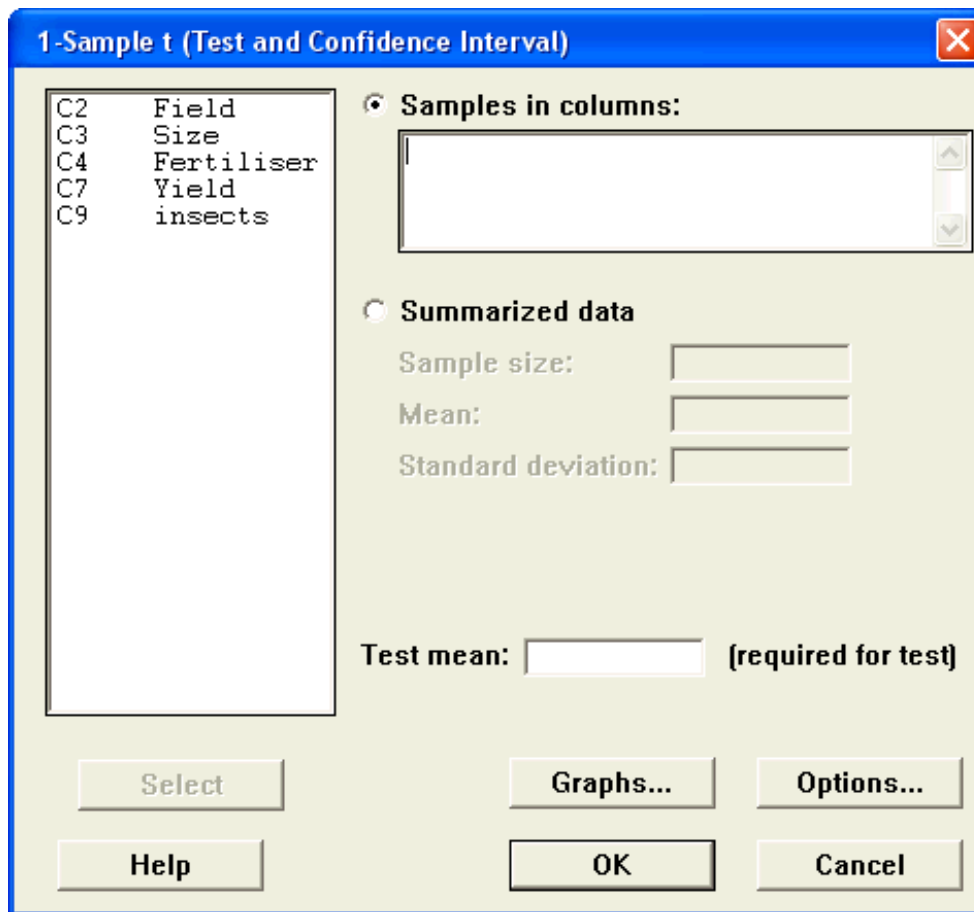


Fig. 7c. Corresponding Minitab dialogue for one-sample t -tests

Almost all statistical packages can read Excel files, so, if you need more statistical facilities than are provided by the combination of Excel and SSC-Stat, it is straightforward to transfer your data.

7.3. Getting the right answer

Some statisticians claim Excel should not be used for statistical work, because the algorithms it uses are numerically unsound. For examples, look at:

McCullough BD and Wilson B (2002) On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. *Computational Statistics and Data Analysis* 40(4), 713-721.

However, we do not agree with the conclusions of the authors that this is sufficient evidence for Excel to be avoided for statistical work. With simple examples, Excel only goes wrong with quite odd data sets, and we think you would notice when they are so. With this proviso, we suggest that the use of Excel is fine for descriptive statistics such as graphs and tables.

Excel includes a statistical toolkit, the Analysis ToolPak, which adds facilities for multiple regression and other types of analysis. Here we are less comfortable: even the Microsoft online support warns about this aspect; see <http://support.microsoft.com/kb/829208/en-us> to read a "Description of the effects of the improved statistical functions for the Analysis ToolPak in Excel 2003".

Hence we suggest that if you need such methods, you should add a statistics package to your repertoire. Others agree with our view, see for example <http://www.npl.co.uk/ssfm/ssfm1/validate/testing/excel.html>