Introduction to SPSS 17

0. Some Basic Ground Rules

In point-and-click mode, SPSS operates in two separate windows – the SPSS Data Editor and the SPSS Viewer. The Data Editor, needless to say, contains the data; each column is a different variable, each row a separate case or individual. It has two modes of viewing – Data View, which resembles a spread sheet (but isn’t), and Variable View, in which variable attributes are set or amended. Switch between the two modes via the tags at the bottom of the Data Editor. All output appears in the SPSS Viewer in a continuous stream of charts, tables and text. You can save the contents of the Data Editor or the Viewer separately; saving either won’t save both.

When you select an SPSS menu item, you will usually be presented with a dialogue box which you must complete before giving the SPSS the OK to proceed. If you don’t know the purpose of any label, box or button which appears in a dialogue box, right click it to display a brief note about its function.

SPSS produces most of its non-graphical output in Pivot Tables. Double click a table to activate the Pivot Table editor. This will either be in a separate window, if the table is overly wide, or in the Viewer window itself. You can change fonts, number of decimal places displayed and even the cell entries themselves if you don’t like them. You can swap the order in which rows or columns appear by dragging the row or column headers. When selecting multiple entries, it is very easy to drag and drop cell entries to a new location, overwriting what is there; so tread carefully. A row of asterisks in a cell indicates the cell is not wide enough to display its contents. Drag the cell border in editing mode to rectify this problem. An SPSS table can be transferred to Microsoft Word, say, by a simple copy and paste, although a copy and paste special as a picture will reproduce the table exactly as it appears in SPSS.

Double clicking graphical output will activate the Chart Editor. See Section 7.2 for more details on operating the Chart Editor.

The SPSS Viewer also contains text output in the form of annotation and syntax. The syntax is the underlying SPSS code which is invoked by your pointing and clicking. This contains a permanent record of exactly what SPSS did in response to your every OK; so you can see what SPSS understood your menu order to be and check that it corresponds to what you thought you were ordering. Syntax can also be copied and pasted to a new or existing SPSS Syntax Editor window, edited and then re-executed. For simple analyses it is usually easier to point and click again or recover the last used dialogue box; but with practice, some efficiencies can be made via the syntax approach. If you find yourself involved in serious data analysis, some expertise in SPSS syntax will be extremely advantageous to you.
1. Transformations

1.1 Transform > Compute Variable...

The SPSS Data Editor resembles a spreadsheet, but its gridded cells lack any computational facility. This deficiency is overcome by the Transform > Compute Variable... procedure, which can create new variables from those already existing.

1.1.1 Type the name of the new variable in the Target Variable: box. No blanks allowed. Certain other symbols are also banned; restrict yourself to letters and numerals to be sure. If you inadvertently type in an existing variable name (or forget to change the previous entry when you next use Transform > Compute Variable), that variable will be replaced in the transformation, but you will be warned before any obliteration takes place. Take heed of the warning; there is no easy way to recover the overwritten values once you’ve pressed OK.

1.1.2 Type the required transformation in the Numeric Expression: box. It isn’t case sensitive and blanks are ignored. Plus and minus are what you would expect. Multiplication is an asterisk (*), division a slash (/) and ** raises to a power. Standard precedence conventions apply to the arithmetical operators; use brackets if in doubt. Oft used functions include Ln, Exp, Sqrt for logarithm (to base e or log_e), exponential function (or inverse log_e) and square root respectively. Calculation takes place case by case; so that the new variable will be populated by the numeric expression applied to every row in turn. If V1 and V2 are existing variable names, then the following are valid numeric expressions:
2 enters the number 2 in all rows of the new variable.

\( V1 \times V2 \) multiplies a value in \( V1 \) by the value in the corresponding row of \( V2 \)

\( V1^{**2} \) squares the values in \( V1 \)

\( 1/V1 \) inverts the values in \( V1 \)

\( \ln(V1) \) takes \( \log_e \) of the value in \( V1 \) (negative \( V1 \) values return a missing value)

\( \ln(V1)^{**2} \) squares the values of \( \log_e \) of \( V1 \)

To expedite the entry of a numeric expression, select from the list of available functions on the right, variables on the left and mathematical operators on the central keypad. Select either All or Arithmetic in the Function group: box to make the list of Functions and Special Variables: appear. Clicking the name of a function will display the function’s action; double clicking it will transfer it to the numeric expression box, as will selecting the name and clicking the up arrow. To transfer an existing variable into a numeric expression, double click it or click it once and then the right arrow. The variable will be inserted at the cursor. No amount of clicking, cajoling or abusive language will transfer an existing variable name into the Target Variable: box.

1.1.3 Click the Type and Label... button only if you want to assign a label to the new variable without further delay. A label offers the opportunity for an extended description of a variable, beyond what might be possible in its single word name. Since labels can be assigned at any time via Variable View in the SPSS Data Editor, this button would appear redundant. However, it does afford a useful option of giving the new variable a label which is the numeric expression itself; so that at some later time you can remember how the variable was obtained.

1.1.4 Upon clicking OK, a new column should appear in the Data Editor window, with a name as supplied by you in the Target Variable: box; either that or an error message.
2. **Analyze > Descriptive Statistics**

This menu item contains an eclectic assortment of procedures for obtaining both numerical and graphical summaries of your data. There is considerable overlap in the capabilities of the procedures on offer and a general lack of cohesion.

2.1 **Analyze > Descriptive Statistics > Frequencies…**

The name suggests that this procedure might only apply to categorical data, but this is not the case. It provides a minimal list of summary statistics and graphical summaries, including histograms, for many variables at a time.

![Frequencies dialog](image)

2.1.1 Place the variables that you want to summarise in the **Variable(s):** box, by selecting (either singly or collectively) from the list of variables on the left and then clicking the right arrow. Tick the **Display frequency tables** box only if your list contains categorical variables or variables which have relatively few values compared to the number of cases. Any truly continuous variable will yield a long table of numeric values all with frequency 1. Such tables can be suppressed under the **Format...** option (see section 2.1.4). Judicious use of this option can reveal anomalies, such as excessive rounding, in data which you might have thought were continuous.

2.1.2 Click the **Statistics...** button to select the numerical summary statistics that you want displayed. The default values here should suffice. The **Values are group midpoints** box appears to be a throwback to pre-calculator days and should be left unchecked.
2.1.3 Click the **Charts...** button if you want to obtain histograms of your variables. If you have categorical variables, you can also draw bar and pie charts, but not at the same time. The histograms are the same as those produced by **Graphs > Legacy Dialogs > Histogram...** (see Section 6.1).

2.1.4 Click the **Format...** button to arrange statistical summary output side by side (**Compare variables**) or sequentially (**Organise output by variables**). The **Order by** options relate to output in any frequency tables produced. If you have ticked the **Display frequency tables** box at step 2.1.1 and you have variables in your list that are not categorical, tick the **Suppress tables** box and set the **Maximum number of categories** to some fraction (say 20%) of the number of cases. This will keep uninformative output to a minimum.
2.2  **Analyze > Descriptive Statistics > Descriptives…**

Provides a less than adequate selection of summary statistics and no graphs. It does, however, provide the option of standardising your variables (subtract the mean and divide by the standard deviation) should you ever have the need to do this; a curious place to keep this facility.

2.2.1  Select the variables to be summarised from the list on the left and transfer them to the **Variable(s):** box by clicking the right arrow. **Options…** are similar to those in section 2.1.2 without the quartiles or percentiles.

2.3  **Analyze > Descriptive Statistics > Explore…**

Provides a comprehensive list of summary statistics, boxplots, histograms and Q-Q plots for many variables at a time, with the option of separate summaries for sub-groups of the data defined by a categorical variable, or factor. This would seem to be all that one could desire. Unfortunately, there is little control over what summary statistics are included in the output, and none at all over the appearance of the output. These features would not be drawbacks if the output appeared in a sensible side by side fashion, for comparative purposes, instead of in a single long list. The graphical output is in **Graphs > Legacy Dialogs** style (see Section 6).
2.3.1 Select the variables for which you want summaries from the list on the left and transfer them to the Dependent List: box by clicking the right arrow. (The term dependent for variable is borrowed from regression analysis.)

2.3.2 If you want separate summaries for subgroups of cases defined by a categorical variable, such as sex, transfer it to the Factor List: box. (Factor is another name for a categorical variable in analysis of variance parlance; strictly speaking for a categorical variable which is an independent variable.) If you leave the Factor List: box empty, no subdivision will take place. If you put more than one categorical variable in the Factor List: box, you will get a separate subdivision and consequent summary for each factor in the list, rather than a single cross factor subdivision. That is, if you were to specify sex and age as factors, you would get a summary breakdown of each (dependent) variable by sex and by age separately rather than a cross tabulated age by sex breakdown. See Section 3 for details on how to achieve the latter.

2.3.3 The Label Cases by: box relates to labelling of outlying points on any boxplots that you draw. By default they are labelled by case number. If you have an alternate case identification variable, you have the option of specifying it here.

2.3.4 If either Both or Statistics is checked under Display, a long list of summary statistics will be output for each (dependent) variable specified. If you want to add extreme value detection, percentiles or M-estimators to the list, click the Statistics… button. The long default list is the Descriptives option under this button. You can suppress this list entirely but not amend it.

2.3.5 Specify the plots that you want to appear in the output via the Plots… button, which will be activated provided either Both or Plots is checked under Display.
Selecting the Dependents together option under Boxplots will plot different variables side by side on the same chart. This is seldom a good idea, unless all variables are measured in identical units or your data is not in standard stacked format and the different variables are actually the same variable measured on different groups. As a rule of thumb, stick with Factor levels together whether or not you have specified any factors at step 2.3.2. This will place the various groups (or single group) defined by your factor (or lack of it) side by side on the same chart, with a separate chart for each (dependent) variable.

Stem-and-leaf plots were designed as an easily constructed alternative to histograms for small data sets, assuming that you didn’t have access to anything other than pencil and paper. Select Histograms in any other circumstance. If you require Q-Q plots, tick Normality plots with tests.

The Spread vs. Level with Levene Test option will only be activated if you have specified a factor at step 2.3.2. It is designed to assist you in the context of a one-way analysis of variance, where you suspect the various group variances are not equal and some form of transformation may be necessary. Selecting None will reduce the amount of time it takes you to get no assistance from the other options. The Power estimation option has nothing whatsoever to do with estimating the power of a test of significance.

2.3.6 The Options... button conceals options on the treatment of missing values. This is a curious but quite common feature of SPSS dialogue boxes. The appearance of an Options... button appears to be mandatory, and when no other options can be thought of, there’s always missing values to fall back on, be the options relevant or not. You need only worry about this if there are missing values in your data and you have specified more than one (dependent) variable at step 2.3.1.

If you Exclude cases listwise, then any case which has a missing value in any one (dependent) variable will be eliminated entirely from consideration. Only cases with values for all (dependent) variables will be included in the summaries. Thus, not all extant data will be summarised. This option is designed for those proceeding to analyses such as multiple regression, where missing values in any one variable create havoc. Strangely, here it appears to be the default.
The **Exclude cases pairwise** option again borrows from multiple regression terminology, rather eccentrically. Under this option, cases with missing values are only eliminated from calculations which could not otherwise proceed. In the context of obtaining summaries of individual variables, this means that a missing value in any one variable only affects calculations for that variable. So, all non-missing data is included in the summaries. This would appear to be the obvious choice for preliminary data exploration.

The **Report Values** option is only applicable if the missing values occur in a categorical variable that you have declared on the **Factor List**. A missing value group is formed in addition to the other groups defined by the factor's non-missing values. The circumstances in which this option might prove useful are difficult to imagine.

### 2.4 Analyze > Descriptive Statistics > Crosstabs…

This procedure is strictly for categorical variables. It provides summary statistics and p-values for the extent of association between two categorical variables. This is essentially a vehicle to perform chi-squared tests of association or equal proportions, but other statistics and tests can be selected. To classify these as descriptive techniques is stretching the definition, so the procedure seems to be misplaced. It does provide multi-way cross tabulation tables for categorical variables, which might just qualify as descriptive.

**Crosstabs**... expects un-aggregated data. That is, every case is a separate individual. If your data is already in the form of frequencies, before you run **Crosstabs**... you must run **Data > Weight Cases**... and **Weight cases by** the variable which contains the cross tabulated frequencies or counts. Remember to return SPSS to individual mode after the event via **Data > Weight Cases**... > **Do not weight cases**.
2.4.1 Select a categorical variable from the list on the left and place it in the **Row(s):** box by clicking the right arrow. Similarly, place your second categorical variable in the **Column(s):** box. This will define a two-way table, with the cells of the table defined by the various cross classifications of the two variables involved. Which variable you choose for the rows and which for the columns is entirely up to you, and your choice will only affect the look of the table, in an obvious way. For a simple Chi-squared test of association or equal proportions on a single two-way table, this is all that is necessary.

If you place more than one categorical variable in either the **Row(s):** or **Column(s):** box, you will get a series of two-way tables, formed from every possible pairwise combination of row and column variables that you have specified. This option only makes sense if your data is in un-aggregated form.

If you place a categorical variable in the **Layer 1 of 1** box, you will get a three-way cross classification, with two-way tables being produced for every category in the variable selected as the first layer. You can continue specifying more categorical variables in subsequent layers, to proceed to higher multi-way cross classifications. Bear in mind that all statistical analyses provided here relate to two-way tables only and so you will have a plethora of p-values to point at if you choose to go down the multi-way path.

2.4.2 By default, the cells of the table contain the observed frequencies only. If you are doing a chi-squared analysis, you might like to include more information to assist in the interpretation of the result. Click the **Cells...** button and choose to display cell expected values, standardised residuals and row or column percentages.

2.4.3 Click the **Statistics...** button. A world of possibilities will be opened up to you. For simplicity, tick **Chi-square.** Tick the **Risk** option if you require confidence intervals for the odds-ratio; another curious choice of nomenclature.

2.4.4 Click **Exact...** if you have small cell frequencies or sparse data. Check the **Exact** box on the dialogue box which opens. Be prepared for a long wait if you have a large number of cross classification categories or large overall sample size. In these circumstances you could consider a Monte Carlo approximation or a reduction in the number of categories by judicious amalgamation.

2.4.5 Tick **Display clustered bar charts** for a graphical summary. The bars are clustered on the column variable.
2.5 Analyze > Descriptive Statistics > Q-Q Plots…

A Q-Q plot can be used to assess whether or not your data follow a Normal distribution (or any other distribution you choose as the norm). The quantiles (i.e. percentiles) of your data are plotted against those of a Normal distribution with the same mean and standard deviation as your data. If the points roughly follow a straight line, your data is Normal. Systematic departures from a straight line indicate skewness (an arc) or long-tailedness (an S shape) in your data. Bear in mind that very few statistical analyses actually require the raw data to follow a Normal distribution. The statistical models assume the “demeaned” data or residuals are Normally distributed.

2.5.1 Transfer the variables whose Normality is under scrutiny to the Variables: box. The default settings of all other options should suffice.
3. **Analyze > Tables > Custom Tables...**

Unfortunately, the simple **Basic Tables...** procedure has been phased out of Version 17. A minimal selection of summary statistics is available for many variables at a time, with the facility to split the data into subgroups defined by the cross classification categories of more than one factor. No graphical summaries are available.

3.1 Construction of a table is best done in **Compact** viewing mode. You can change to **Normal** viewing mode at any time. Drag the variables to be summarised from the **Variables:** list on the left to the table construction area. This is best done as a multiple selection, but can be done individually. Where you place the variables in the construction area will determine whether they will form the rows or columns of the table.

3.2 If you want to break the data into subgroups according to the values of some categorical variable or variables, drag the categorical variables to the vicinity of the **Rows** or **Columns** bar, or for a layering effect click **Layers:** and a new box will open. Where precisely you drag the variables to will determine the look of the table. Click **Normal** to see if the result matches your expectations.

3.3 The **Summary Statistics...** button controls what summary statistics appear in the table. Select a statistic/s of your choice from the **Statistics:** box and transfer it to the **Display:** box by clicking the arrow. Multiple selections are
possible via shift or control click. Change the number format and number of decimals displayed as required. Click **Apply to All** if you want the chosen statistics to be calculated for all displayed variables, or **Apply to Selection** if you want different statistics for different variables. In the latter case, it appears that you can only select one variable at a time and you must do this before entering the **Summary Statistics...** dialogue box.

3.4 The **Position:** setting in the **Summary Statistics** box (no, not the **Summary Statistics** dialogue box but the **Summary Statistics** area on the original **Custom Tables** dialogue box) controls how the summary statistics appear in the tables, assuming you have selected more than one. That is either across the page or down the page.

3.5 The **Categories and Totals...** button opens another can of worms, allowing you to specify whether you want row or column margins to be included in the table, and if so, what you want in them.

3.6 Click the **Test Statistics** tab at the top left of the **Custom Tables** box if you dare. Chi-squared tests are available, if all you have is count data, or pairwise t-tests otherwise; with or without Bonferroni correction and sensible variance estimation.
4. **Analyze > Compare Means**

The procedures under the **Compare Means** umbrella have mostly been superseded by the more comprehensive ones under **General Linear Model**. They have limited flexibility, but that makes them comparatively easy to operate. **Once you have mastered these procedures, you should move on.** They do not encourage an holistic approach to data analysis.

4.1 **Analyze > Compare Means > One-Sample T Test…**

4.1.1 Put the variable or variables you want to test in the **Test Variable(s):** box by selecting from the list on the left and clicking the right arrow. A separate analysis will be done for each variable listed. Set the **Test Value:** to the hypothesised value of the population mean. **Options…** is the usual missing values standby (see Section 2.3.6). Click **OK**.

4.2 **Analyze > Compare Means > Independent-Samples T Test…**

This procedure assumes your data is in standard stacked format, so typically there will be a categorical variable or factor which divides the cases into two or more groups. The groups can only be tested for equality of means two at a time; this is a two (independent) samples t test. The categorical variable can be defined during the procedure.
4.2.1 Put the variable or variables of interest in the **Test Variable(s):** box, by making the selection from the list on the left and clicking the right arrow. A separate analysis will be done for each variable listed.

4.2.2 Put the categorical variable which defines the separate groups in the **Grouping Variable:** box by selecting it and clicking the right arrow. Since the process allows the existence of more than two categories in your grouping variable, you must specify precisely which two your attention is currently focused on by clicking the **Define Groups...** button. This will be case even if you only have two categories to choose from.

![Define Groups](image)

The values that you enter for **Group1:** and **Group 2:** must be values that appear in your grouping variable for the process to proceed smoothly. The process will test whether the difference in group 1 mean – group 2 mean is 0.

If you don’t have a categorical variable, you have the option of defining it now, by putting any numeric variable in the **Grouping Variable:** box and specifying a **Cut point:** on that variable, thereby dividing the cases into two groups.

4.2.3 The **Options...** button only deals with missing values (see Section 2.3.6) and confidence interval percentages, so click **OK.**

4.3 **Analyze > Compare Means > Paired-Samples T Test...**

This procedure assumes the paired observations are measured on the same individual, and so belong to the same case. Thus it expects the data to be in two separate variables, unstacked style. It tests whether the mean difference between the pairs is 0.
4.3.1 Select the two paired variables in the list on the left either by shift or control clicking. Then transfer them to the Paired Variables: box by clicking the arrow. You can also transfer the variables one at a time if you prefer. The difference is taken as Variable1 - Variable2. If this isn’t convenient click the two headed swap arrow on the right.

You can select as many pairs as you care to. A separate analysis will be done for each pair of variables listed. The up and down arrows on the right change the order in which the pairs are analysed.

4.3.3 The Options... button deals with missing values (see Section 2.3.6) and confidence interval percentages, so click OK.

4.4 Analyze > Compare Means > One-Way ANOVA

4.4.1 Put the variable or variables of interest in the Dependent List: by selecting from the list on the left and clicking the right arrow. A separate analysis will be done for each variable specified.

4.4.2 Transfer the categorical variable (one only) which defines the groups to the Factor: box, by the selection and right arrow method.

4.4.3 The Post Hoc... option hides a multitude of sins or multiple comparison techniques, which amounts to the same thing. Select LSD, unless your journal editor strongly (and erroneously) insists you use Tukey, Bonferroni and Scheffe are statistically correct and hopelessly conservative and the rest tend to make unwarranted assumptions about the structure of your groups. Always interpret multiple comparison results with a grain of salt and never in the absence of corroborative evidence from the ANOVA that some difference plausibly exists. If you don’t have equal variances, you shouldn’t be here.
4.4.4 Click the Options... button and at least check Descriptive. This will give you a table of means and standard deviations. The confidence intervals are based on the individual group standard errors and should be discounted. Check Homogeneity of variance test only if you have any faith in Levene. Use it a guide only. The Means plot is usually unedifying and should only be attempted if your category codes actually represent values on a numeric (or perhaps ordinal) scale. Leave Fixed and random effects to a later date and just leave the rest.

4.4.5 The Contrasts... option provides an alternative to Post Hoc multiple comparisons. It’s a little too sophisticated for a first course. Wait until you have done polynomial regression and know more about general linear models.
5. **Analyze > Regression > Linear…**

Variables are entered into boxes by clicking first their name in the left hand list and then the appropriate right arrow.

5.1 Enter the y variable into the **Dependent**: box. This is the response variable.

5.2 Enter the x variable into the **Independent(s)**: box.

5.3 Click the **Save** button. Choose **Unstandardized Predicted Values** and **Studentized Residuals** from the smorgasbord (possibly also **Leverage values** - these are SPSS patented “centred” leverages - and **Cook’s Distances** if you know what to do with them). If you require confidence and prediction intervals select **Mean** and **Individual Prediction Intervals**. Anything you select from the available alternatives will be added as new columns to your data sheet. This can make for extremely messy housekeeping, particularly as these settings are remembered from regression to regression. Keep a close eye on which columns belong to which regression and cull as necessary.

5.4 The other buttons either provide acceptable default settings or don’t include any options that are relevant. Click **OK**.
6. **Histograms and Boxplots**

In SPSS Version 17 there are three methods of drawing graphs; the old old way under **Graphs > Legacy Dialogs**, the old new way under **Graphs > Legacy Dialogs > Interactive** and the new way under **Graphs > Chart Builder**. The newer ways afford the opportunity of drawing rotatable bivariate histograms, and no longer appear to leave points off the plot in certain circumstances. The old old way is by no means inferior to the newer ways when drawing similar plots. Neither the old nor the newer methods allow you to draw simple histograms and boxplots for a list of variables, which is where the graphical summary procedures under **Analyze > Descriptive Statistics** have the advantage.

### 6.1 **Graphs > Legacy Dialogs > Histogram…**

![](image)

6.1.1 Place the variable (yes, only one) of interest in the **Variable:** box, by selecting it from the list on the left and clicking the right arrow.

6.1.2 If you want to break your variable down into subgroups according to the values of one or more categorical variable, then place these variables in the **Panel by** boxes. Histograms will be panelled across and/or down the page according to whether you specify the categorical variable in **Columns:** and/or **Rows:**. Multiple variables can be specified in the **Panel by** boxes, with a choice of nesting or crossing (the default) the category combinations.

6.1.3 Click **OK**.

6.1.4 Double clicking the histogram in the **SPSS Viewer** will open the **Chart Editor**. Double click any element of the chart in the editor to make changes to that element. For example, double click the histogram bins to open the bin
Properties dialogue. Under the Binning tag you can change the bin width or number of bins and the starting point (anchor) of the first bin. Change the all important colour of the bins under the Fill & Border tag.

6.2 Graphs > Legacy Dialogs > Interactive > Histogram…
6.2.1 Drag the variable to be graphed from the list on the left to the x-axis position. For a relative frequency histogram, as opposed to a frequency histogram, drag the system variable \texttt{Percent($pct$)} to the y-axis to replace \texttt{Count($count$)}.

6.2.2 If you want to break the data into subgroups, drag the grouping categorical variable or variables to the Panel Variables box. There appears to be less flexibility in the layout than in the Legacy Dialog method.

6.2.3 Click the Histogram tab to specify the interval (bin) width before the event. It can also be changed after the event. Adjusting the Start Point under this tag will cause you nothing but grief. Leave it alone.

6.2.4. The plot when drawn looks exactly the same as that produced by Graphs > Histogram... in Section 6.1 with half the effort required here. Double click the plot to open the same Chart Editor as in Section 6.1.4.

6.3 Using Graphs > Chart Builder... to Construct a Histogram

6.3.1 Click on Histogram in the bottom left hand Choose from: list in the Gallery of charts that have been prepared earlier.

6.3.2 Double click the simple histogram icon (left most of the four) in the Gallery box. A prototype histogram should appear in the chart preview area above the Gallery. This histogram does not now nor will it ever bear any relationship to the histogram which will finally be drawn. In addition the histogram Element Properties box should open. If it doesn’t you can always click the Element Properties... button middle right.
6.3.3 Drag the variable (just the one) of interest from the left hand Variables: box to the x-axis box immediately beneath the histogram. You can adjust some aspects of the histogram at this point in the Element Properties window, or do it post production in the Chart Editor. If you want a relative frequency histogram, specify it now in the Element Properties window; make sure Bar1 is highlighted in the Edit Properties of: box and then select Histogram Percent in the Statistic: box in the Statistics area. It’s too late once the histogram is drawn.

6.3.4 You can break the data into subgroups by clicking the Groups/Points ID tag and selecting either Rows or Columns panel variable and then dragging your categorical variable to the new box which opens in the chart preview area. There seems to be limited flexibility here.

6.3.5 Finally click OK. Once the histogram is drawn, double clicking it opens the same Chart Editor as in 6.1.4 and 6.2.4.

6.4 Graphs > Legacy Dialogs > Boxplot…

There are four (count them) different dialogue box entry points to this labyrinthine procedure guarded by yet another impenetrable dialogue box. Tukey would turn in his grave. Some simple examples are explained below.

6.4.1 For a simple, single variable boxplot select Simple and check Summaries of separate variables before clicking Define. Transfer your variable to the Boxes Represent: box. If you transfer more than one variable, all will appear on the same plot. This is not a good idea, unless all are measured on the same scale. Extreme values are labelled by case number by default, or whatever identifying variable you place in the Label cases by: box. You can divide the data into subgroups by specifying Panel by: categorical variables, but this is best done by parallel boxplots or clustered boxplots. Once drawn the plot can be edited in the Chart Editor. Double click the plot to activate it.
6.4.2 For a simple, parallel boxplot, select **Simple** and check **Summaries for groups of cases** before clicking **Define**. Transfer your variable to the **Variable**: box. Transfer the categorical variable which defines the subgroups to the **Category Axis**: box. The other boxes are the standard fare. Editing after the event is done in the **Chart Editor**, opened by double clicking the plot.
6.5 **Graphs > Legacy Dialogs > Interactive > Boxplot…**

This method has a simpler entry strategy than that of Section 6.4.

6.5.1 For a simple, single variable boxplot, drag the variable from its place in the left hand list to the vertical axis.

6.5.2 For a simple parallel boxplot, drag the variable to the vertical axis and a group-defining categorical variable to the horizontal axis.
6.5.3 You can cluster a second subgrouping by dragging another categorical variable to either of the **Legend Variables** boxes; **Color:** or **Style:** (for the colour or patterning distinguishing the boxes of the subgroups, of course). Alternatively, you can go for a 3-D effect.

6.5.4 A further or alternative display of subgroupings is possible via the **Panel Variables** box.

6.6 **Using Graphs > Chart Builder… to Construct a Boxplot**
6.6.1 Click on **Boxplot** in the bottom left hand **Choose from:** list in the **Gallery** of charts that have been prepared earlier.

6.6.2 For a single boxplot, double click the 1-D boxplot icon (right most of the three) in the **Gallery** box. A prototype boxplot should appear in the chart preview area above the **Gallery**. This boxplot does not bear much resemblance to the boxplot which will finally be drawn. In addition the boxplot **Element Properties** box should open. If it doesn’t you can always click the **Element Properties…** button middle right.

6.6.3 Drag the variable (just the one) of interest from the left hand **Variables:** box to the y-axis box immediately to the left of the prototype boxplot. You can adjust some aspects of the boxplot at this point in the **Element Properties** box, or do it post production in the **Chart Editor**.

6.6.4 For a simple parallel boxplot, double click the simple boxplot icon (left most of the three) in the **Gallery** box. Drag the variable of interest to the y-axis and the categorical variable to the x-axis. Somewhat confusingly, if you do not populate the x-axis with any variable, then a 1-D boxplot will be drawn.

6.6.5 More complicated subgrouping arrangements are available under the **Groups/Point ID** tag.

6.6.5 If you crave a 3-D effect on your boxplot, don’t use **Graphs > Chart Builder**.
7. Scatter Plots

As with histograms and boxplots there are three ways of producing exactly the same plot: **Graphs > Chart Builder...** (there’s actually two methods inside the **Builder**, via the **Gallery** tag or the **Basic Elements** tag), **Graphs > Legacy Dialogs > Scatter/Dot...** and **Graphs > Legacy Dialogs > Interactive > Scatterplot...**. Given that **Legacy Dialogs** might disappear in the near future, it might be worth familiarising yourself with the **Chart Builder**. However, in the short term, the original **Graphs > Legacy Dialogs > Scatter/Dot** remains as simple and quick to operate as any of the other procedures and can draw a wider range of plots. The **Chart Builder** also has a tendency to be a trifle flaky, particularly in **Basic Elements** mode.

7.1 **Graphs > Legacy Dialogs > Scatter/Dot**

The first thing you will see is the following Dialogue box:

Choose **Simple Scatter** if you have one y and one x (possibly with subgroups defined by a third factor and panelled by a fourth or more factors), **Overlay Scatter** if you have many y-x pairs to plot on the same graph and **Matrix Scatter** if you want a plot of all possible pairwise combinations from a list of variables. Having made your choice, click the **Define** button and further dialogue boxes will open. When the plot is drawn, you can add fit lines, interpolate between points, change the default markers and much more via the **Chart Editor**. Note, it does not appear possible to construct an overlay plot with the **Chart Builder** or via **Legacy Dialogs > Interactive**.

7.1.1 **Simple Scatter**

Place your y variable in the **Y Axis**: box by selecting it in the list of variables on the left and clicking the right arrow. Similarly, place your x variable in the **X Axis**: box. If you want to distinguish the points on the plot according to subgroups defined by the levels of a third factor, place that factor in the **Set Markers by**: box. The plot markers (i.e. the symbols used for the points on the plot) will vary according to which subgroup the point belongs to.

Labelling cases attaches a label to every point on the plot. The labels are taken from the variable you specify in the **Label Cases by**: box. Typically they are text labels. A labelled plot is a very busy plot. The information is in the points, not the labels. If you have more than one factor defining subgroups, you can break the plot into separate panels according to the cross classification levels of as many factors as you care to.
7.1.2 Overlay Scatter

Select two variables from the list on the left (shift or control click) and transfer them to the Y-X Pairs: box by clicking the right arrow, or by dragging them there. You can also select and transfer the variables singly. Which variable ends up as the Y Variable and which the X Variable depends on the method of...
selection. If it is the wrong way round, select the pair in the **Y-X Pairs:** box and click the two headed swap arrow on the right to interchange the variables. You can specify as many pairs of variables as you want to appear on the plot, but the plot may look very strange if the ranges of the x variables and/or the ranges of the y variables are even mildly different.

The **Label Cases by:** and **Panel by** variables have the same function as in the **Simple Scatter** plot.

7.1.3 Matrix Scatter

![Matrix Scatter Diagram](image)

Either select multiple variables from the list on the left (shift or ctrl click) and transfer them to **Matrix Variables:** box via the right arrow, or drag them one at a time to this box. The result will be a two-dimensional array of scatter plots, in which every matrix variable will be plotted against every other matrix variable. (See example on next page.) The axes can’t be labelled with anything other than the variable name and plot definition diminishes with the number of variables, but if you want to avoid endless incarnations of **Simple Scatter** to get a quick look at your data, this is the way to do it.

The **Set Markers by:**, **Label Cases by:** and **Panel by** variables have the same function as previously.
<table>
<thead>
<tr>
<th>Case</th>
<th>Tonnage</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
7.2 Chart Editor

The Chart Editor is invoked by double clicking any plot in the SPSS output window, irrespective of which Graphs procedure produced the plot. The Chart Editor opens in a separate window. Closing it will return you to the output window with any chart modifications you have made intact.

Double clicking any element of the plot will open an edit box for that element. Double click items on the legend (if you have one) to edit individual data subgroups (defined by the Set Markers by: variable). Alternatively, select items to edit from the Chart Editor menus. You can add up to cubic polynomials to the data set as a whole or to subgroups of the data. You can also interpolate between points to join the dots (useful to connect fitted values) or type in an equation of a line to add to the plot. If you don’t like the look of SPSS’s default markers, you can save your own choices via File > Save Chart Template and invoke the template either when you draw the plot or after the event in the Chart Editor. This could save you countless hours of needless frustration.