Chapter 4: Exploring data with graphs

Self-test answers

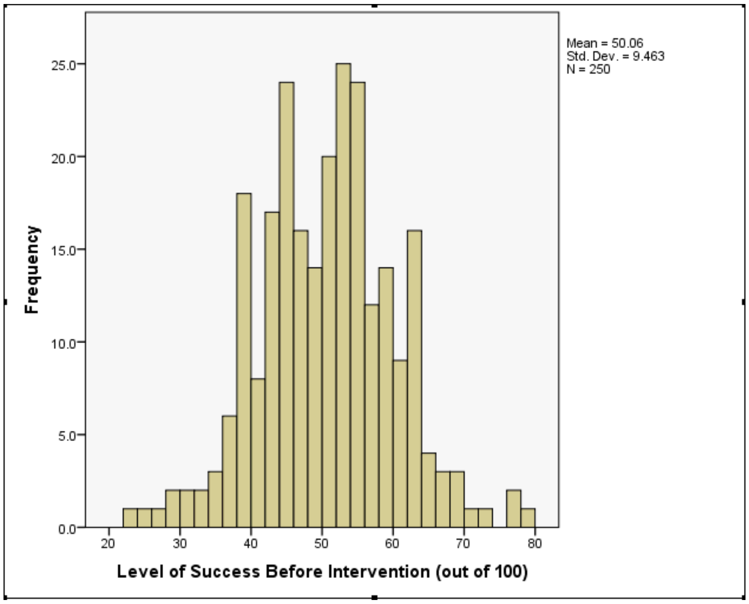
SELF-TEST What does a histogram show?

A histogram is a graph in which values of observations are plotted on the horizontal axis, and the frequency with which each value occurs in the data set is plotted on the vertical axis.

SELF-TEST Produce a histogram and population pyramid for the success scores *before* the intervention.

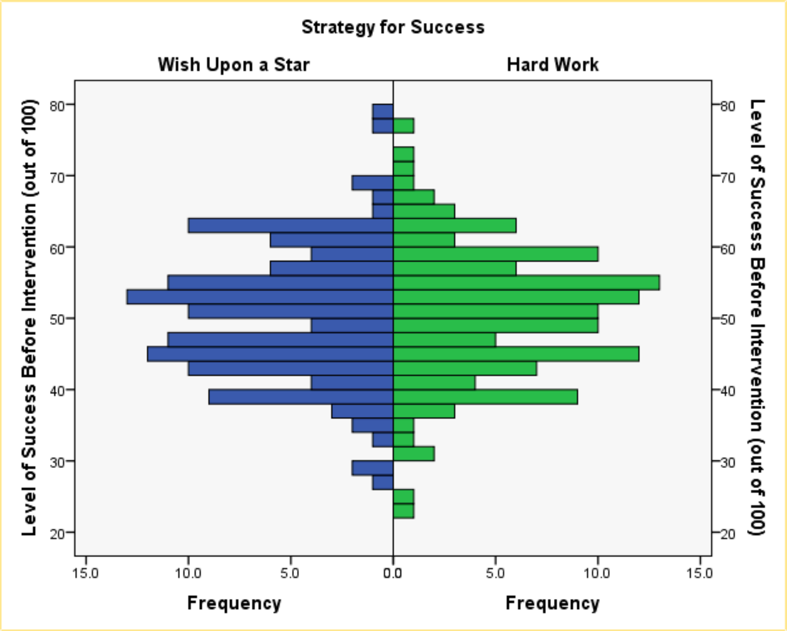
First, access the Chart Builder and then select *Histogram* in the list labelled *Choose from* to bring up the gallery. This gallery has four icons representing different types of histogram, and you should select the appropriate one either by double-clicking on it, or by dragging it onto the canvas in the Chart Builder. We are going to do a simple histogram first, so double-click on the icon for a simple histogram. The *Chart Builder* dialog box will show a preview of the graph in the canvas area. Next, click on the variable (**Success_Pre**) in the list and drag it to X-Axis?. You will now find the histogram previewed on the canvas. (Although SPSS calls the resulting graph a preview, it's not really because it does not use your data to generate this image — it is a preview only of the general form of the graph, and not what your specific graph will actually look like.) To draw the histogram click on OK.

The resulting histogram is shown below. Looking at the histogram, the data look fairly symmetrical and there doesn't seem to be any sign of skew.



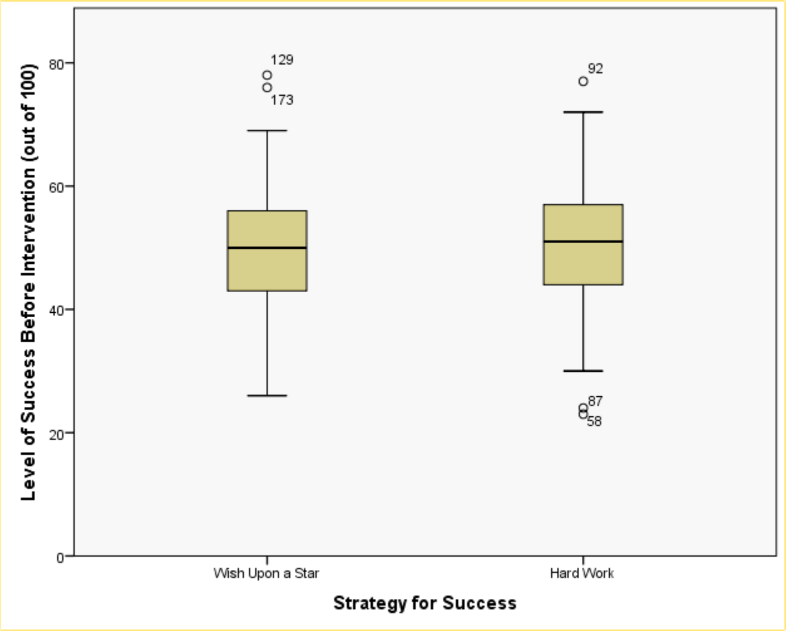
To compare frequency distributions of several groups simultaneously we can use a population pyramid. Click on the population pyramid icon (see the book chapter) to display the template for this graph on the canvas. Then from the variable list select the variable representing the success scores before the intervention and drag it into Distribution Variable? to set it as the variable that you want to plot. Then select the variable **Strategy** and drag it to Split Variable? to set it as the variable for which you want to plot different distributions. Click on OK to produce the graph.

The resulting population pyramid is show below and looks fairly symmetrical. This indicates that both groups had a similar spread of scores before the intervention. Hopefully, this example shows how a population pyramid can be a very good way to visualise differences in distributions in different groups (or populations).



SELF-TEST Produce boxplots for the success scores *before* the intervention.

To make a boxplot of the pre-intervention success scores for our two groups, double-click on the *simple boxplot* icon, then from the variable list select the **Success_Pre** variable and drag it into Y-Axis? and select the variable **Strategy** and drag it to X-Axis?. Note that the variable names are displayed in the drop zones, and the canvas now displays a preview of our graph (e.g. there are two boxplots representing each gender). Click on OK to produce the graph.



Looking at the resulting boxplots above, notice that there is a tinted box, which represents the IQR (i.e., the middle 50% of scores). It's clear that the middle 50% of scores are more or less the same for both groups. Within the boxes, there is a thick horizontal line, which shows the median. The workers had a very slightly higher median than the wishers, indicating marginally greater pre-intervention success but only marginally.

In terms of the success scores, we can see that the range of scores was very similar for both the workers and the wishers, but the workers contained slightly higher levels of success than the wishers. Like histograms, boxplots also tell us whether the distribution is symmetrical or skewed. If the whiskers are the same length then the distribution is symmetrical (the range of the top and bottom 25% of scores is the same); however, if the top or bottom whisker is much longer than the opposite whisker then the distribution is asymmetrical (the range of the top and bottom 25% of scores is different). The scores from both groups look symmetrical because the two whiskers are similar lengths in both groups.

SELF-TEST The procedure for producing line graphs is basically the same as for bar charts except that you get lines on your graphs instead of bars. Therefore, you should be able to follow the previous sections for bar charts but selecting a simple line chart instead of a simple bar chart, and selecting a multiple line chart instead of a clustered bar chart. I would like you to produce line charts of each of the bar charts in the previous section. In case you get stuck, the self-test answers that can be downloaded from the companion website will take you through it step by step.

Simple Line Charts for Independent Means

To begin with, imagine that a film company director was interested in whether there was really such a thing as a 'chick flick' (a film that typically appeals to women more than men). He took 20 men and 20 women and showed half of each sample a film that was supposed to be a 'chick flick' (*Bridget Jones's Diary*), and the other half of each sample a film that didn't fall into the category of 'chick flick' (*Memento*, a brilliant film by the way). In all cases he measured their arousal as a measure of how much they enjoyed the film. The data are in a file called **ChickFlick.sav** on the companion website. Load this file now.

Let's just plot the mean rating of the two films. We have just one grouping variable (the film) and one outcome (the arousal); therefore, we want a simple line chart. Therefore, in the Chart Builder double-click on the icon for a simple line chart. On the canvas you will see a graph and two drop zones: one for the *y*-axis and one for the *x*-axis. The *y*-axis needs to be the dependent variable, or the thing you've measured, or more simply the thing for which you want to display the mean. In this case it would be **arousal**, so select arousal from the variable list and drag it into the *y*-axis drop zone (Y-Axis?). The *x*-axis should be the variable by which we want to split the arousal data. To plot the means for the two films, select the variable **film** from the variable list and drag it into the drop zone for the *x*-axis (X-Axis?).

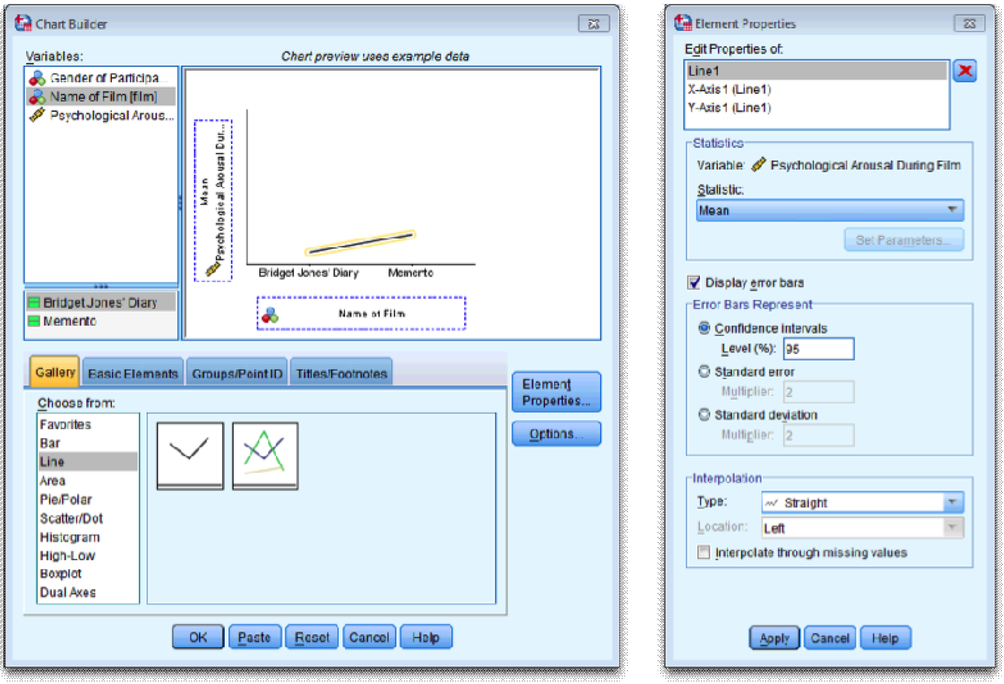


Chart Builder

Variables:
- Gender of Participa...
- Name of Film [film]
- Psychological Arous...

- Bridget Jones' Diary
- Memento

Chart preview uses example data

Mean Psychological Arousal Dur...

Bridget Jones' Diary   Memento

Name of Film

Choose from: Line

Element Properties

Edit Properties of: Line1

Statistics

Variable: Psychological Arousal During Film

Statistic: Mean

☑ Display error bars

Error Bars Represent

● Confidence intervals   Level (%): 95

○ Standard error   Multiplier: 2

○ Standard deviation   Multiplier: 2

Interpolation

Type: Straight

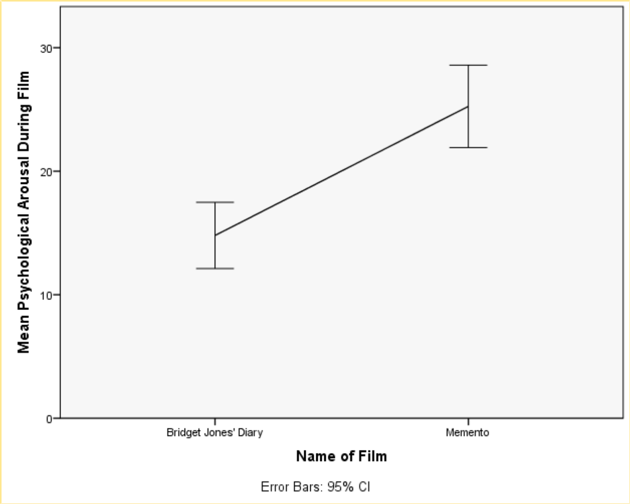
Location: Left

□ Interpolate through missing values

Dialog boxes for a simple line chart with error bars

The figure above shows some other options for the line chart. The main dialog box should appear when you select the type of graph you want, but if it doesn't click on

Element Properties... in the Chart Builder. There are three important features of this dialog box. The first is that, by default, the lines will display the mean value. This is fine, but just note that you can plot other summary statistics such as the median or mode. Second, you can adjust the form of the line that you plot. The default is a straight line, but you can have others like a spline (curved line). Finally, we can ask SPSS to add error bars to our line chart by selecting ☑ Display error bars. We have a choice of what our error bars represent. Normally, error bars show the 95% confidence interval, and I have selected this option (● Confidence intervals). Note, though, that you can change the width of the confidence interval displayed by changing the '95' to a different value. You can also display the standard error (the default is to show 2 standard errors, but you can change this to 1) or standard deviation (again, the default is 2, but this could be changed to 1 or another value). It's important that when you change these properties you click on Apply: if you don't then the changes will not be applied to Chart Builder. Click on OK to produce the graph.



Line chart of the mean arousal for each of the two films

The resulting line chart displays the means (and the confidence interval of those means). This graph shows us that on average, people were more aroused by *Memento* than they were by *Bridget Jones's Diary*. However, we originally wanted to look for gender effects, so this graph isn't really telling us what we need to know. The graph we need is a *multiple line graph*.

Multiple line charts for independent means

To do a multiple line chart for means that are independent (i.e., have come from different groups) we need to double-click on the multiple line chart icon in the Chart Builder (see the book chapter). On the canvas you will see a graph as with the simple line chart but there is now an extra drop zone: Set color. All we need to do is to drag our second grouping variable into this drop zone. As with the previous example, select **arousal** from the variable list and drag it into Y-Axis?, then select **film** from the variable list and drag it into X-Axis?. In addition, though, we can now select the **gender** variable and drag it into Set color. This will mean that lines representing males and females will be displayed in different colours. As in the previous section, select error bars in the properties dialog box and click on Apply to apply them to the Chart Builder. Click on OK to produce the graph.

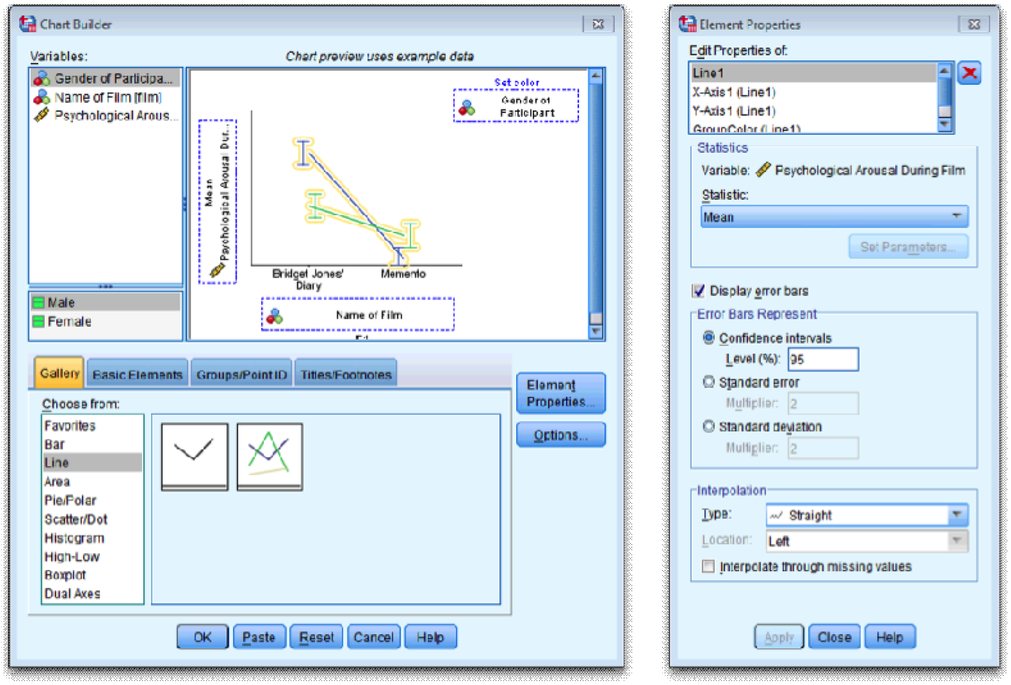


Chart Builder

Variables: Gender of Participa..., Name of Film [film], Psychological Arous...

Element Properties

Edit Properties of: Line1, X-Axis1 (Line1), Y-Axis1 (Line1), GroupColor (Line1)

Statistics

Variable: Psychological Arousal During Film

Statistic: Mean

☑ Display error bars

Error Bars Represent

● Confidence intervals Level (%): 95

○ Standard error Multiplier: 2

○ Standard deviation Multiplier: 2

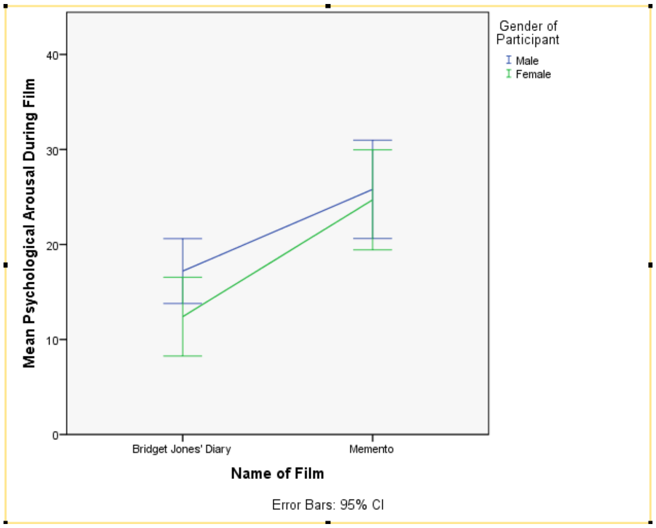
Interpolation

Type: Straight

Location: Left

☐ Interpolate through missing values

Dialog boxes for a multiple line chart with error bars



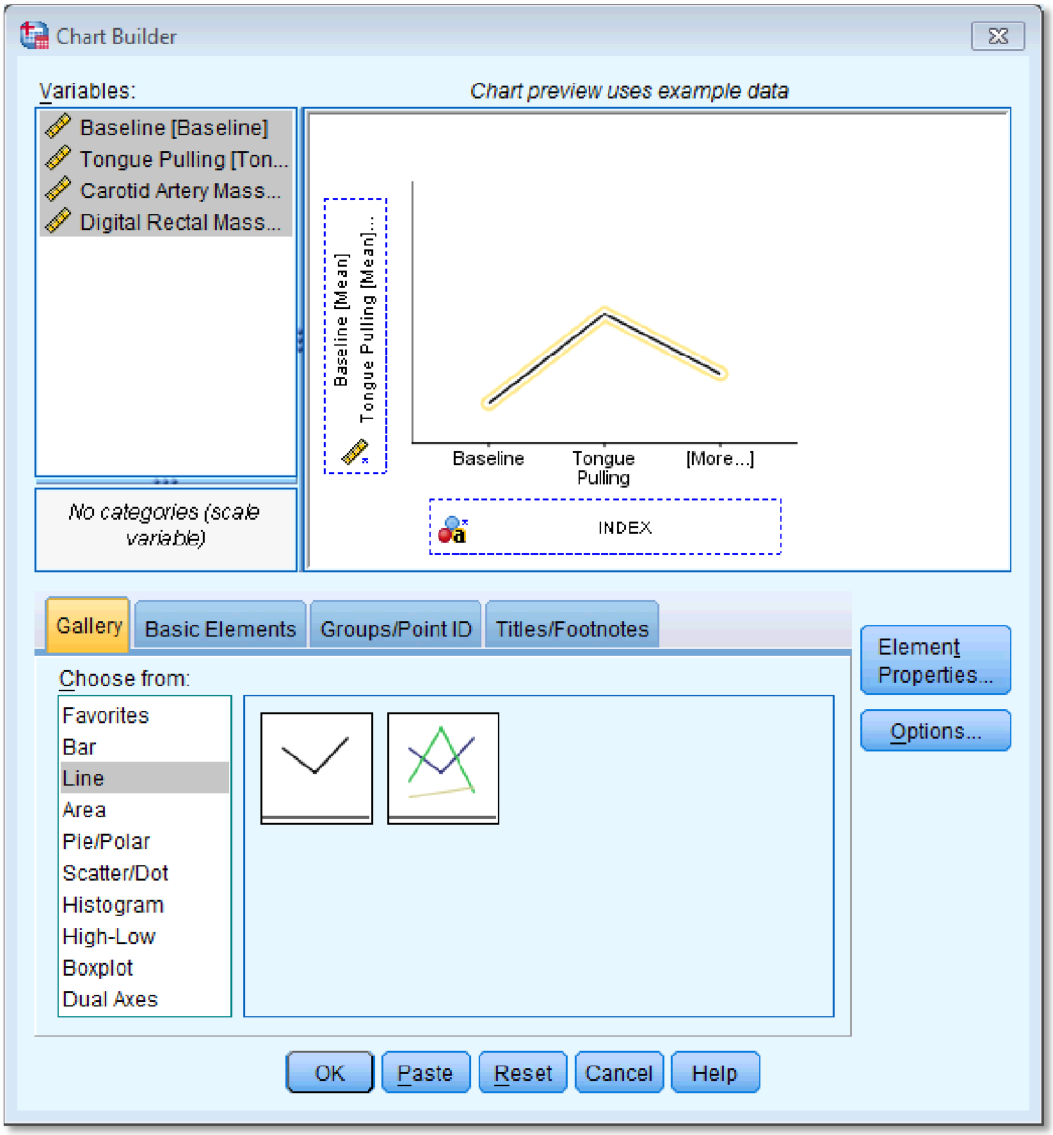
Line chart of the mean arousal for each of the two films.

The resulting line chart tells us the same as the simple line graph: that is, arousal was overall higher for *Memento* than for *Bridget Jones's Diary*, but it also splits this information by gender. Look first at the mean arousal for *Bridget Jones's Diary*; this shows that males were actually more aroused during this film than females. This indicates they enjoyed the film more than the women did! Contrast this with *Memento*, for which arousal levels are comparable in males and females. On the face of it, this contradicts the idea of a 'chick flick': it actually seems that men enjoy chick flicks more than the chicks (probably because it's the only help we get to understand the complex workings of the female mind!).

Simple line charts for related means

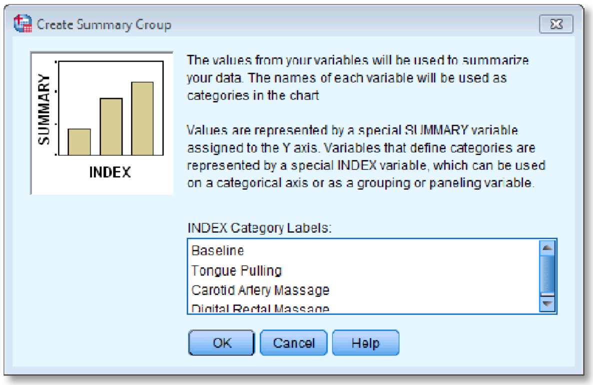
Hiccups can be a serious problem. Charles Osborne apparently got a case of hiccups while slaughtering a hog (well, who wouldn't?) that lasted 67 years. People have many methods for stopping hiccups (a surprise, holding your breath), and medical science has put its collective mind to the task too. The official treatment methods include tongue-pulling manoeuvres, massage of the carotid artery and, believe it or not, digital rectal massage (Fesmire, 1988). I don't know the details of what the digital rectal massage involved, but I can probably imagine. Let's say we wanted to put this to the test. We took 15 hiccup sufferers, and during a bout of hiccups administered each of the three procedures (in random order and at intervals of 5 minutes) after taking a baseline of how many hiccups they had per minute. We counted the number of hiccups in the minute after each procedure. Load the file **Hiccups.sav**. Note that these data are laid out in different columns; there is no grouping variable that specifies the interventions because each patient experienced all interventions. In the previous two examples we have used grouping variables to specify aspects of the graph (e.g., we used the grouping variable **film** to specify the *x*-axis). For repeated-measures data we will not have these grouping variables and so the process of building a graph is a little more complicated (but only a little).

To plot the mean number of hiccups go to the Chart Builder and double-click on the icon for a simple line chart. As before, you will see a graph on the canvas with drop zones for the *x*- and *y*-axes. Previously we specified the column in our data that contained data from out outcome measure on the *y*-axis, but for these data we have four columns containing data on the number of hiccups (the outcome variable). What we have to do then is to drag all four of these variables from the variable list into the *y-axis* drop zone. We have to do this simultaneously. First, we need to select multiple items in the variable list. Select the first variable by clicking on it with the mouse. The variable will be highlighted in yellow. Now, hold down the *Ctrl* key on the keyboard and click on a second variable. Both variables are now highlighted in yellow. Again, hold down the *Ctrl* key and click on a third variable in the variable list and so on for the fourth. In cases in which you want to select a list of consecutive variables, you can do this very quickly by simply clicking on the first variable that you want to select (in this case **baseline**), then hold down the *Shift* key on the keyboard and then click on the last variable that you want to select (in this case **digital rectal massage**); notice that all of the variables in between have been selected too. Once the four variables are selected you can drag them by clicking on any one of the variables and then dragging them into Y-Axis? as shown in the figure:



Specifying a simple line chart for repeated-measures data

Once you have dragged the four variables onto the *y*-axis drop zones a new dialog box appears. This box tells us that SPSS is creating two temporary variables. One is called **Summary**, which is going to be the outcome variable (i.e., what we measured – in this case the number of hiccups per minute). The other is called **Index**, which will represent our independent variable (i.e., what we manipulated – in this case the type of intervention). SPSS uses these temporary names because it doesn't know what our particular variables represent, but we should change them to something more helpful. Just click on OK to get rid of this dialog box.



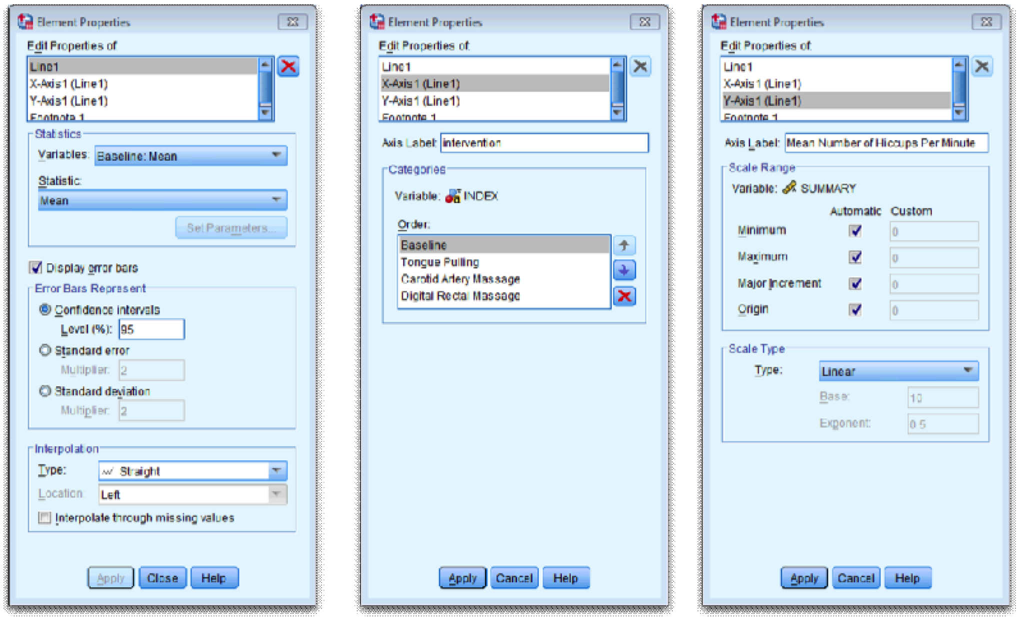
Create Summary Group

The values from your variables will be used to summarize your data. The names of each variable will be used as categories in the chart.

Values are represented by a special SUMMARY variable assigned to the Y axis. Variables that define categories are represented by a special INDEX variable, which can be used on a categorical axis or as a grouping or paneling variable.

INDEX Category Labels:
Baseline
Tongue Pulling
Carotid Artery Massage
Digital Rectal Massage

The *Create Summary Group* dialog box



Element Properties

Edit Properties of: Line1 (selected); X-Axis1 (Line1); Y-Axis1 (Line1); Footnote 1

Statistics
Variables: Baseline: Mean
Statistic: Mean

☑ Display error bars

Error Bars Represent
● Confidence intervals — Level (%): 95
○ Standard error — Multiplier: 2
○ Standard deviation — Multiplier: 2

Interpolation
Type: Straight
Location: Left
□ Interpolate through missing values

Element Properties

Edit Properties of: Line1; X-Axis1 (Line1) (selected); Y-Axis1 (Line1); Footnote 1

Axis Label: intervention

Categories
Variable: INDEX
Order:
Baseline
Tongue Pulling
Carotid Artery Massage
Digital Rectal Massage

Element Properties

Edit Properties of: Line1; X-Axis1 (Line1); Y-Axis1 (Line1) (selected); Footnote 1

Axis Label: Mean Number of Hiccups Per Minute

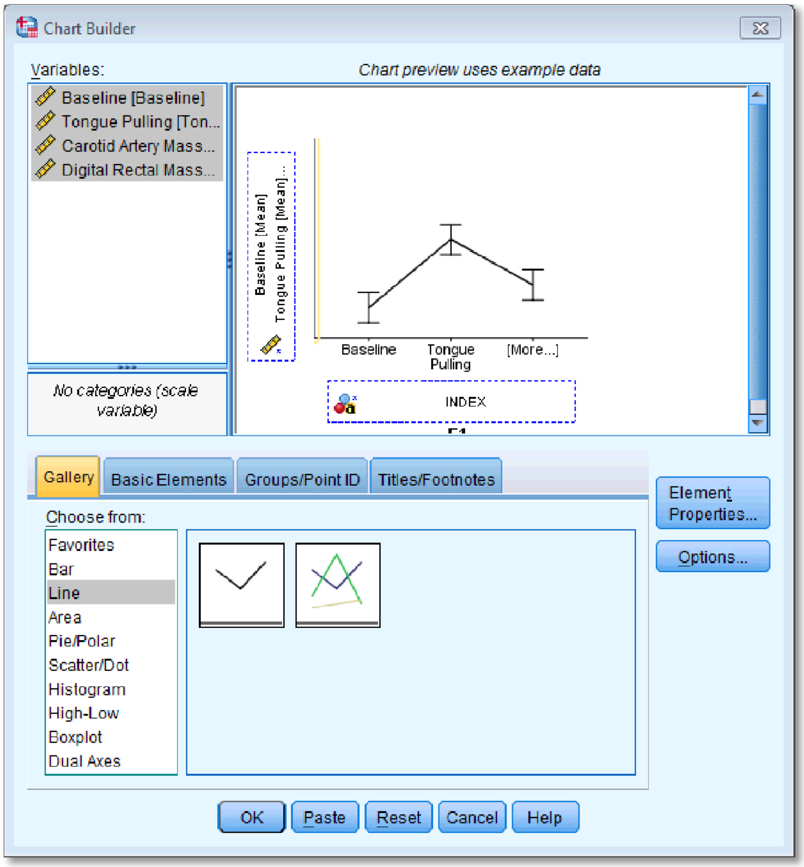
Scale Range
Variable: SUMMARY

	Automatic	Custom
Minimum	☑	0
Maximum	☑	0
Major Increment	☑	0
Origin	☑	0

Scale Type
Type: Linear
Base: 10
Exponent: 0.5

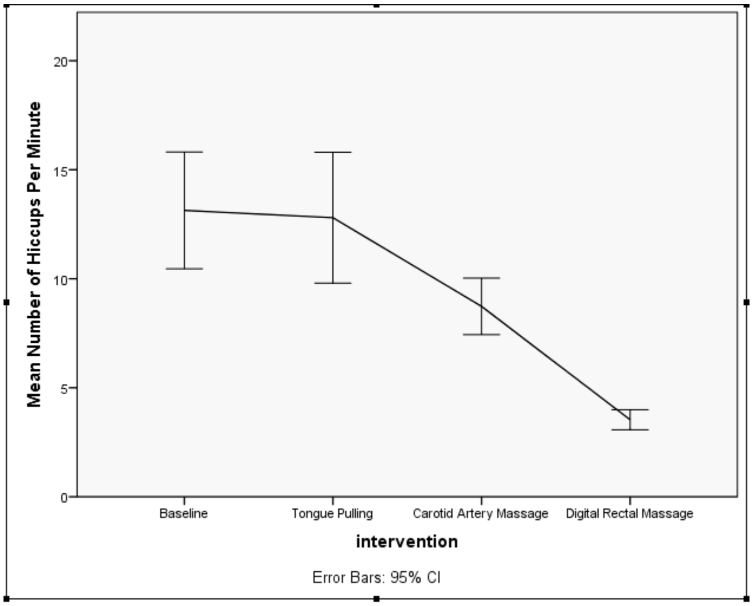
Setting *Element Properties* for a repeated-measures graph

We need to edit some of the properties of the graph. The figure shows the options that need to be set: if you can't see this dialog box then click on Element Properties... in the Chart Builder. In the left panel of the figure just note that I have selected to display error bars (see the previous two sections for more information). The middle panel is accessed by clicking on *X-Axis1 (Line1)* in the list labelled *Edit Properties of* and this allows us to edit properties of the horizontal axis. The first thing we need to do is give the axis a title and I have typed *Intervention* in the space labelled *Axis Label*. This label will appear on the graph. Also, we can change the order of our variables if we want to by selecting a variable in the list labelled *Order* and moving it up down using ↑ and ↓. If we change our mind about displaying one of our variables then we can also remove it from the list by selecting it and clicking on ✕. Click on Apply for these changes to take effect. The right panel is accessed by clicking on *Y-Axis1 (Line1)* in the list labelled *Edit Properties of* and it allows us to edit properties of the vertical axis. The main change that I have made here is to give the axis a label so that the final graph has a useful description on the axis (by default it will just display 'Mean', which isn't very helpful). I have typed 'Mean Number of Hiccups Per Minute' in the box labelled *Axis Label*. Also note that you can use this dialog box to set the scale of the vertical axis (the minimum value, maximum value and the major increment, which is how often a mark is made on the axis). Mostly you can let SPSS construct the scale automatically and it will be fairly sensible – and even if it's not you can edit it later. Click on Apply to apply the changes.



Completed Chart Builder for a repeated-measures graph

Click on OK to produce the graph. The resulting line chart displays the mean (and the confidence interval of the mean) number of hiccups at baseline and after the three interventions. Note that the axis labels that we typed in have appeared on the graph. We can conclude that the amount of hiccups after tongue pulling was about the same as at baseline; however, carotid artery massage reduced hiccups, but not by as much as a good old-fashioned digital rectal massage. The moral here is: if you have hiccups, find something digital and go amuse yourself for a few minutes.



Line chart of the mean number of hiccups at baseline and after various interventions

Multiple line charts for related means

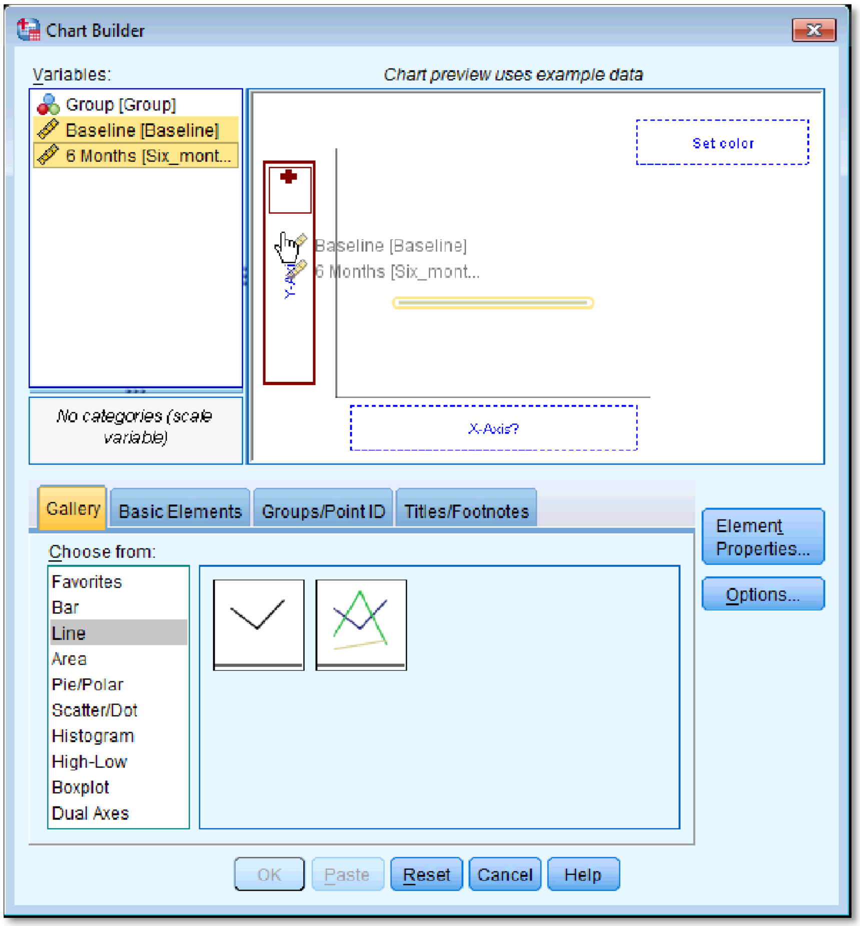
Just like bar charts, these, to the best of my knowledge, can't be done. I could be wrong, though – I often am.

Multiple line charts for 'mixed' designs

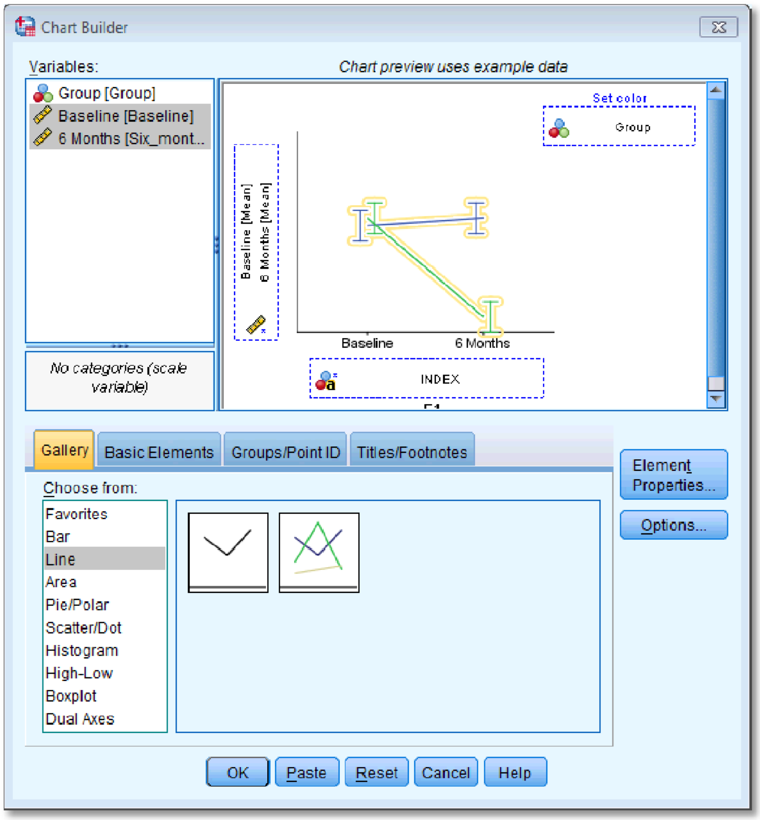
The Chart Builder might not be able to do charts for multiple repeated-measures variables, but it can graph what is known as a mixed design. This is a design in which you have one or more independent variables measured using different groups, and one or more independent variables measured using the same sample. Basically, the Chart Builder can produce a graph provided you have only one repeated-measure variable.

My students like to text-message during my lectures (I assume they text-message the person next to them to say, 'Bloody hell, this guy is so boring I need to poke out my own eyes'). What will happen to future generations, though? Not only will they develop super-sized thumbs, but they might not learn correct written English. Imagine we conducted an experiment in which a group of 25 children was encouraged to send text messages on their mobile phones over a six-month period. A second group of 25 was forbidden from sending text messages for the same period. To ensure that kids in this latter group didn't use their phones, this group were given armbands that administered painful shocks in the presence of microwaves (like those emitted from phones). The outcome was a score on a grammatical test (as a percentage) that was measured both before and after the intervention. The first independent variable was, therefore, text message use (text messagers versus controls) and the second independent variable was the time at which grammatical ability was assessed (baseline or after 6 months). The data are in the file **Text Messages.sav**.

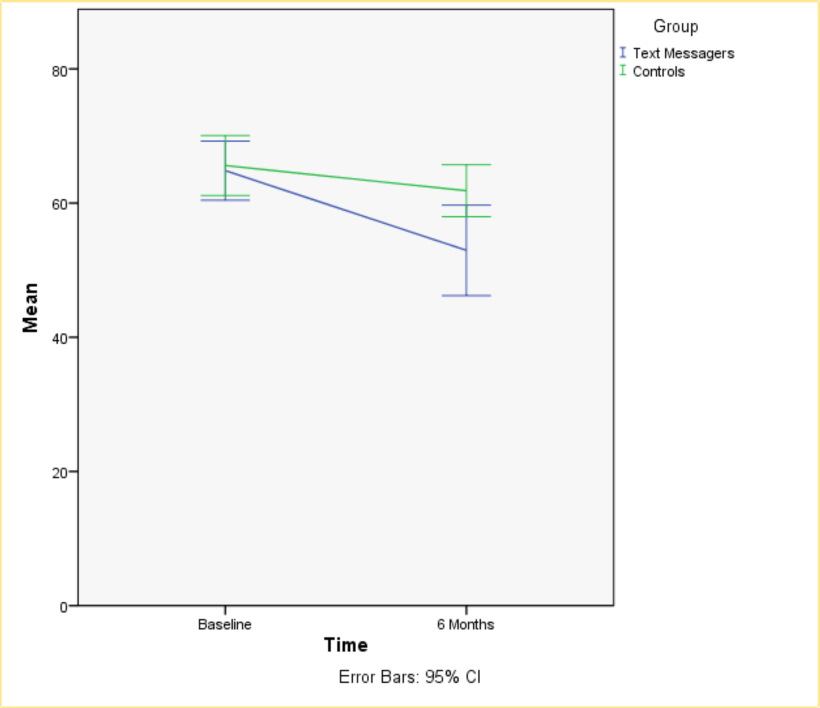
To graph these data we need to follow the general procedure for graphing related means. Our repeated-measures variable is time (whether grammar ability was measured at baseline or 6 months) and is represented in the data file by two columns, one for the baseline data and the other for the follow-up data. In the Chart Builder you need to select these two variables simultaneously by clicking on one and then holding down the *Ctrl* key on the keyboard and clicking on the other. When they are both highlighted, click on either one and drag it into Y-Axis?. The second variable (whether children text-messaged or not) was measured using different children and so is represented in the data file by a grouping variable (**group**). This variable can be selected in the variable list and dragged into Set color. The two groups will now be displayed as different-coloured lines. The finished Chart Builder is below. Click on OK to produce the graph.



Selecting the repeated-measures variable in the Chart Builder



Completed dialog box for an error bar graph of a mixed design



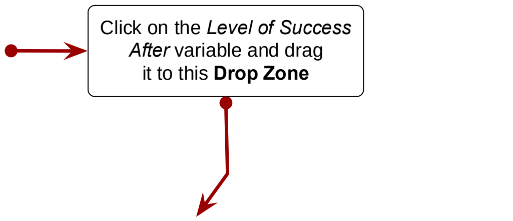
Error bar graph of the mean grammar score over 6 months in children who were allowed to text-message versus those who were forbidden

The resulting line chart shows that at baseline (before the intervention) the grammar scores were comparable in our two groups; however, after the intervention, the grammar scores were lower in the text messagers than in the controls. Also, if you compare the blue line with the green line you can see that text messagers' grammar scores have fallen over the 6 months, whereas the controls' grammar scores are fairly stable over time. We could, therefore, conclude that text messaging has a detrimental effect on children's understanding of English grammar and civilization will crumble, with Abaddon rising cackling from his bottomless pit to claim our wretched souls. Maybe.

SELF-TEST Doing a simple dot plot in the Chart Builder is quite similar to drawing a histogram. Reload the **Jiminy Cricket.sav** data and see if you can produce a simple dot plot of the success scores after the intervention.

First, make sure that you have loaded the **Jiminy Cricket.sav** file and that you open the Chart Builder from this data file. Once you have accessed the Chart Builder (see the book chapter) select the *Scatter/Dot* in the chart gallery and then double-click on the icon for a simple dot plot (again, see the book chapter if you're unsure of what icon to click). The *Chart Builder* dialog box will now show a preview of the graph in the canvas area. At the moment it's not very exciting because we haven't told SPSS which variables we want to plot. Note that the variables in the data editor are listed on the left-hand side of the Chart Builder, and any of these variables can be dragged into any of the spaces surrounded by blue dotted lines (called *drop zones*).

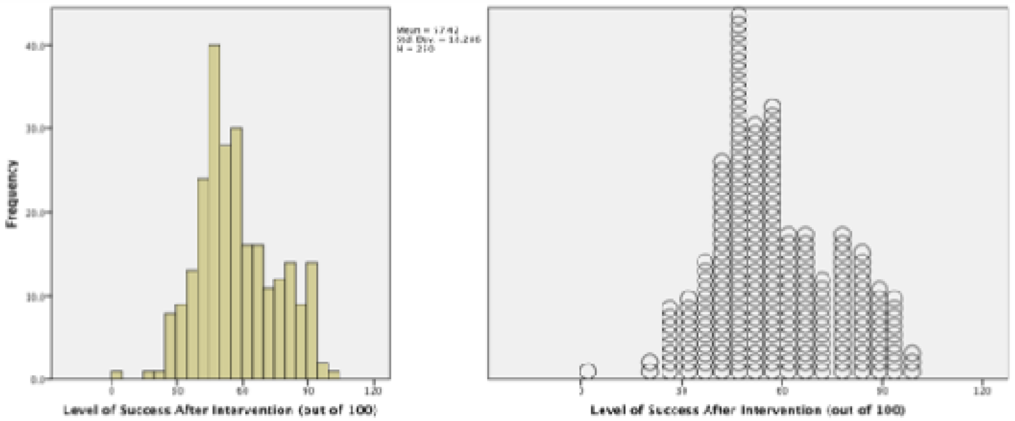
Like a histogram, a simple dot plot plots a single variable (*x*-axis) against the frequency of scores (*y*-axis), so there is just one drop zone (X-Axis?). All we need to do is select a variable from the list and drag it into X-Axis?. To do a simple dot plot of the success scores after the intervention we click on this variable in the variable list and drag it to X-Axis? as shown below; you will now find the dot plot previewed on the canvas. To draw the dot plot click on OK.



Click on the *Level of Success After* variable and drag it to this **Drop Zone**

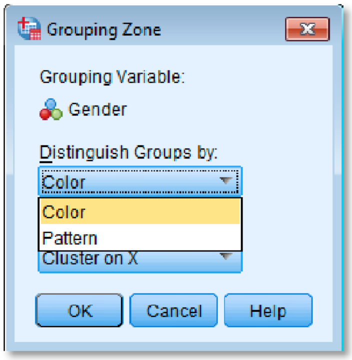
Defining a simple dot plot (a.k.a. density plot) in the Chart Builder

The resulting density plot is shown below along with the original histogram from the book. The first thing that should leap out at you is that they are very similar (in terms of what they show): they both tell us about the distribution of scores, and they both show us the outlier that was discussed in the chapter. These graphs, therefore, are really just two ways of showing the same thing. The density plot gives us a little more detail than the histogram, but essentially they show the same thing.

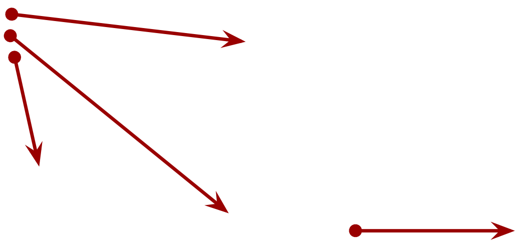


Density plot of the Download day 1 hygiene scores and the original histogram from the book

SELF-TEST Doing a drop-line plot in the Chart Builder is quite similar to drawing a clustered bar chart. Reload the **ChickFlick.sav** data and see if you can produce a drop-line plot of the arousal scores. Compare the resulting graph with the earlier clustered bar chart of the same data.

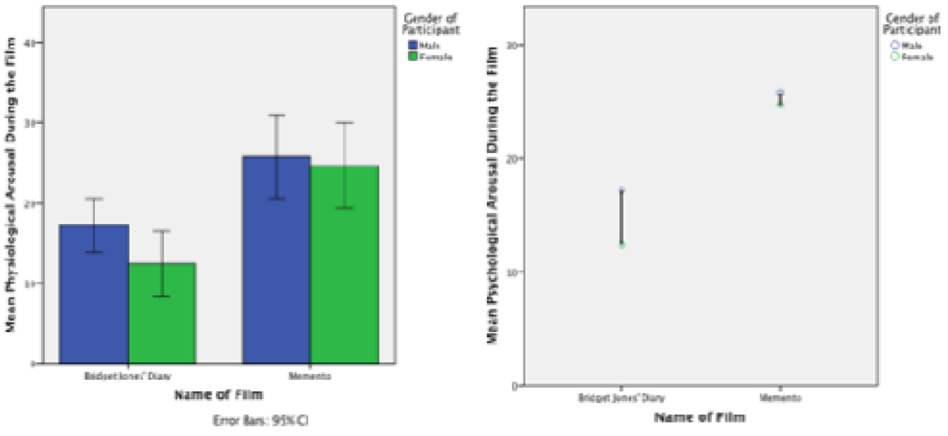


To do a drop-line chart for means that are independent (i.e., have come from different groups) we need to double-click on the drop-line chart icon in the Chart Builder (see the book chapter if you're not sure what this icon looks like or how to access the Chart Builder). On the canvas you will see a graph with some dots and three drop zones that are the same as for a clustered bar chart: X-Axis?, Y-Axis?, and Set color. As with the clustered bar chart example from the book, select **arousal** from the variable list and drag it into Y-Axis?, select **film** from the variable list and drag it into X-Axis?, and select the **gender** variable and drag it into the Set color drop zone. This will mean that the dots representing males and females will be displayed in different colours, but if you want them displayed as different symbols then, to make this change, double-click in the Set color drop zone to bring up a new dialog box. Within this dialog box there is a drop-down list labelled *Distinguish Groups by* and in this list you can select *Color* or *Pattern*. To change the default, select *Pattern* and then click on OK to make the change. Obviously you can switch back to displaying different groups in different colours in the same way. The completed Chart Builder is shown below; click on OK to produce the graph.



Using the Chart Builder to plot a drop-line graph

The resulting drop-line graph is shown below together with the clustered bar chart from the book. Hopefully it's clear that these graphs show the same information (although notice that the *y*-axis has been scaled differently by SPSS so that the differences between films look bigger on the drop-line graph than on the bar chart). In both graphs we can see that arousal was overall higher for *Memento* than for *Bridget Jones's Diary*, that men and women differed very little in their arousal during *Memento*, and that men were more aroused during *Bridget Jones's Diary*. The fact that arousal in males and females differed more for *Bridget Jones's Diary* than for *Memento* is possibly a little clearer in the drop-line graph than the bar chart, but it's really down to preference.

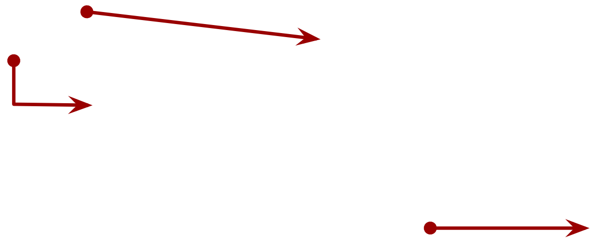


Drop-line graph of mean arousal scores during two films for men and women and the original clustered bar chart from the book

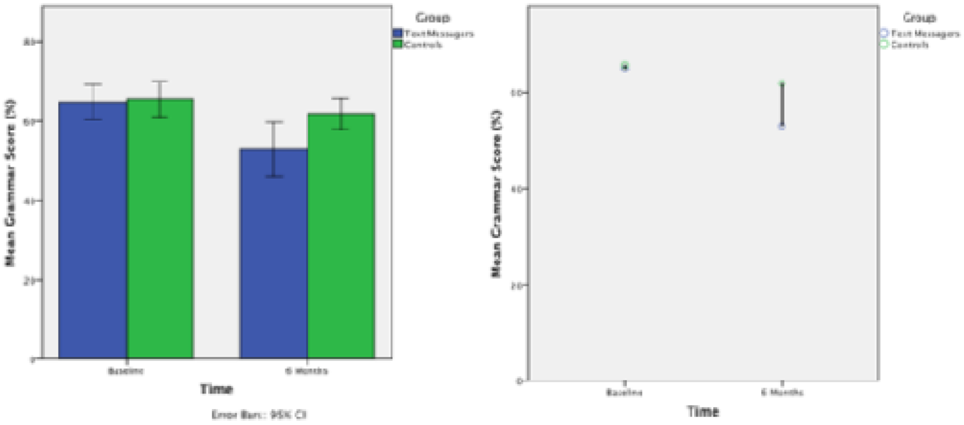
SELF-TEST Now see if you can produce a drop-line plot of the **Text Messages.sav** data from earlier in this chapter. Compare the resulting graph with the earlier clustered bar chart of the same data.

To do a drop-line graph of these data we need to follow the general procedure for graphing related means. First, in the Chart Builder you need to double-click on the icon for a drop-line graph (see the book chapter for help with this if you need it). Our repeated-measures variable is time (whether grammar ability was measured at baseline or 6 months) and is represented in the data file by two columns, one for the baseline data and the other for the follow-up data. Then select these two variables simultaneously by clicking on one and then holding down the *Ctrl* key on the keyboard and clicking on the other. When they are both highlighted, click on either one and drag it into Y-Axis? as shown below. The second variable (whether children text-messaged or not) was measured using different children and so is represented in the data file by a grouping variable (**group**). This variable can be selected in the variable list and dragged into Set color. The two groups will now be displayed as different-coloured dots. The finished Chart Builder is shown below. Click on OK to produce the graph.

The resulting drop-line graph is shown together with the bar chart from the book chapter. They both show that at baseline (before the intervention) the grammar scores were comparable in our two groups. On the drop-line graph this is particularly apparent because the two dots merge into one (you can't see the drop line because the means are so similar). After the intervention, the grammar scores were lower in the text messagers than in the controls. By comparing the two vertical lines it's clearer on the drop-line graph that the difference between text messagers and controls is bigger at 6 months than it is pre-intervention.



Completing the dialog box for a drop-line graph of a mixed design



Error bar graph of the mean grammar score over 6 months in children who were allowed to text-message versus those who were forbidden