

2

Data display, summary and manipulation

In God we trust; all others must bring data! (Attributed to W Edwards Deming)

Overview

Data are essential in the monitoring and improvement of processes and in the measure and control phases of Six Sigma projects. Such data are often obtained in time sequence. For example, a critical dimension might be measured every hour on each member of a sample of machined automotive components during production in a factory, the number of mortgage agreements completed successfully each day might be recorded by a building society. A run chart of such data can frequently be highly informative and forms the basis for some control charts. The construction of run charts using Minitab will be used to introduce the reader to the software and the key features of sessions, projects, worksheets, menus, dialog boxes, graphs, and ReportPad™, etc. The facility for calculation of derived data will also be introduced.

The use of histograms for the display of data will be described, and widely used summary statistics that indicate location and variability defined. The chapter concludes with consideration of a variety of methods for data entry in Minitab, of data manipulation and of the detection of missing and erroneous data values.

2.1 The run chart – a first Minitab session

2.1.1 Input of data via keyboard and creation of a run chart in Minitab

In their book *Building Continual Improvement*, Wheeler and Poling (1998, p. 31) introduce run charts as follows:

Data tend to be associated with time. When was this value obtained? What time period does that value represent? Because of this association with time there is information contained in the time-order sequence of the values. To recover this information you will need to look at your data in a time-ordered plot, which is commonly called a running record or time-series graph.

In order to introduce running records or run charts, consider the time series of weights (g) of glass bottles in Table 2.1. Bottle weight is a key process output variable in the food packaging industry. Each bottle was formed in the same mould of the machine used to produce the bottles and the time interval between sampling of bottles was 15 minutes. The target weight is 490 g and the production run was scheduled to run for a total of 12 hours.

On opening Minitab the screen displayed in Figure 2.1 will appear. Two main windows are visible. The Session window displays the results of analyses in text format; initially it displays the date, time and a message of welcome. (It is also possible to perform tasks by entering commands in the Session window instead of using the Minitab menus.) The Data window

Table 2.1 Initial bottle weight data.

Sample	Weight	Sample	Weight	Sample	Weight	Sample	Weight	Sample	Weight
1	488.1	6	493.1	11	490.5	16	489.7	21	490.2
2	493.4	7	487.4	12	492.2	17	488.5	22	489.8
3	488.7	8	488.4	13	490.6	18	493.6	23	486.1
4	484.4	9	488.6	14	490.8	19	489.1	24	487.0
5	491.8	10	485.9	15	486.7	20	489.4	25	485.4

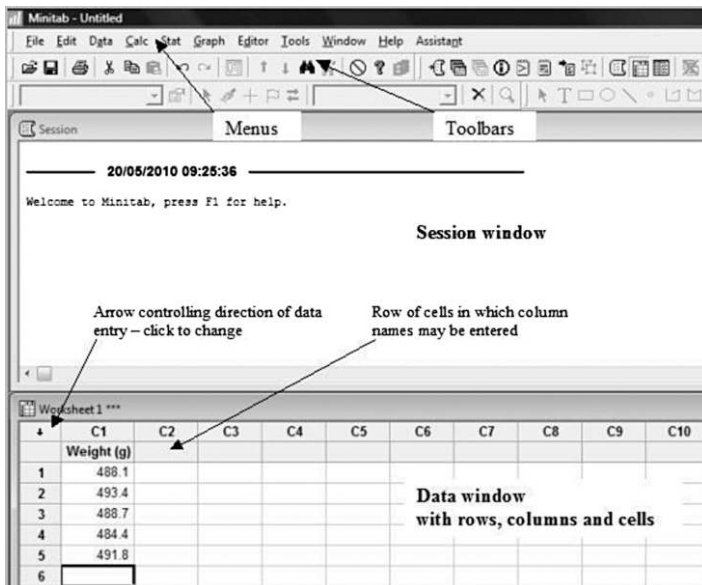


Figure 2.1 Initial Minitab screen.



Figure 2.2 Project Manager toolbar.

displays an open worksheet that has the appearance of a spreadsheet. (It is possible to use multiple worksheets within a single project.) Note that the blue band across the top of the Session window is a deeper colour than that across the top of the Data window, indicating that the Session window is currently active. (Take a moment to click on the blue bands to make the worksheet active and then the Session window active again.) A third component, in addition to the Session and Data windows, of the new Minitab project just opened up, is the Project Manager, which is minimized at this stage. Note the corresponding icon labelled **Proje...** at the foot of the screen.

Figure 2.2 shows the Project Manager toolbar, which has 12 icons, with the mouse pointer located on the icon for the **ReportPad** on the toolbar. Clicking on an icon makes the corresponding component active. Note that the message displayed indicates the project component associated with the icon and gives the keys that may be used to make the component active as an alternative to clicking on its icon. From the extreme left the icons displayed are: Show Session Folder, Show Worksheets Folder, Show Graphs Folder, Show Info, Show History, Show ReportPad, Show Related Documents, Show Design, Session Window, Data Window, Project Manager and Close All Graphs, respectively.

Enter all 25 of the weight data values in Table 2.1 into the first column, C1, of the worksheet in the current Data window, and enter Weight (g) as the column heading to name the variable. Figure 2.1 displays the worksheet with the variable name and the first five data values entered.

In order to access the dialog box for the creation of a run chart first click the **Stat** menu icon (see Figure 2.3) then **Quality Tools** and finally **Run Chart...** (Throughout this book the shorthand **Stat > Quality Tools > Run Chart...** will be used to indicate such sequences of steps.)

The Run Chart dialog box can now be completed as shown in Figure 2.4. Highlight **Weight (g)** in the window on the left of the dialog box and click on the **Select** button so that **Weight (g)** appears in the window labelled **Single column:**. (Alternatively highlight

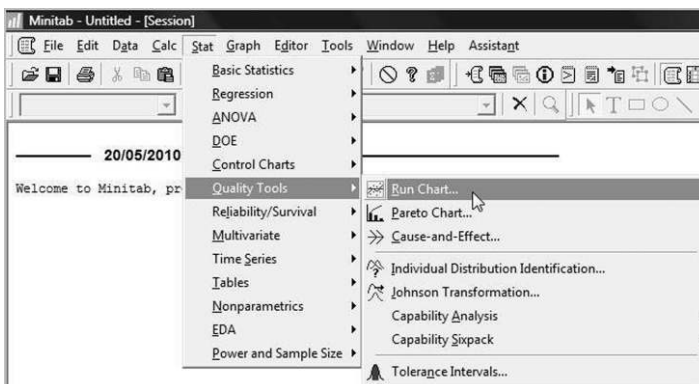


Figure 2.3 Accessing the run chart dialog box.

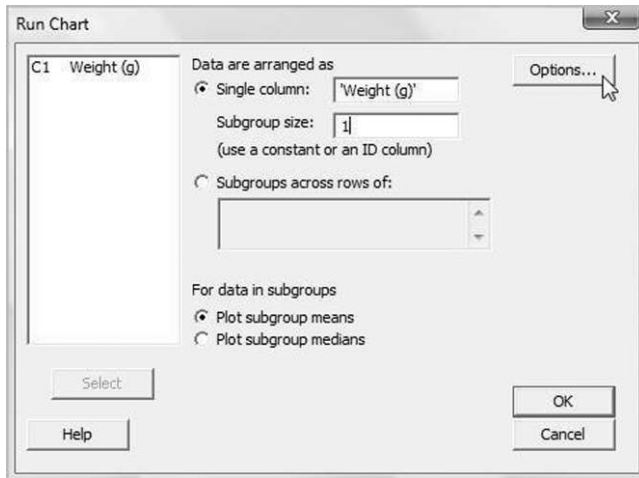


Figure 2.4 Completion of the run chart dialog.

Weight (g) and double-click.) Enter **Subgroup size: 1** in the appropriate window as each sample consisted of a single bottle. (Had the weights of samples of four bottles been recorded every 15 minutes then **Subgroup size: 4** would have been entered.)

Clicking the **Options...** button reveals a subdialog box that may be used to create a title for the run chart as indicated in Figure 2.5. Click **OK** to return to the main dialog box and then click **OK** again to display the run chart – see Figure 2.6. (Had the option to create a title not been used then Minitab would have assigned the title ‘Run Chart of Weight (g)’).

Those involved in running the process can learn about its performance from scrutiny of the run chart. Weight is plotted on the vertical axis, with the weight of each bottle represented by a square symbol and the symbols connected by dotted lines indicating the sequence of sampling. (Had, for example, four bottles been weighed every 15 minutes then Minitab offers the choice of a run chart with symbols corresponding to either the mean or median of the weights of each sample or subgroup of four bottles.) The horizontal axis is labelled Observation – each sample of a single bottle may be thought of as an observation of the process behaviour. The horizontal line on the chart corresponds to the median weight of the entire group of 25 bottles weighed. On moving the mouse pointer to the line a textbox containing the text ‘Reference line at 489.1’ is displayed. The median weight is 489.1 g, which in this case is the weight of bottle number 19.

In the technical language of statistics the median is a measure of location, which gives an indication of ‘where one is at’ in terms of process performance. Scrutiny of the run chart reveals 12 points above the median line, 12 points below the median line and the point

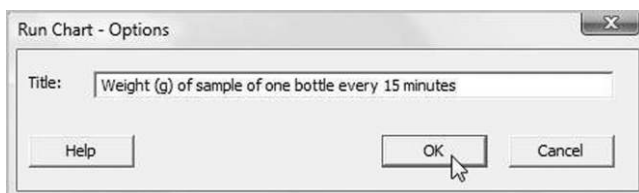


Figure 2.5 Creating a run chart title.

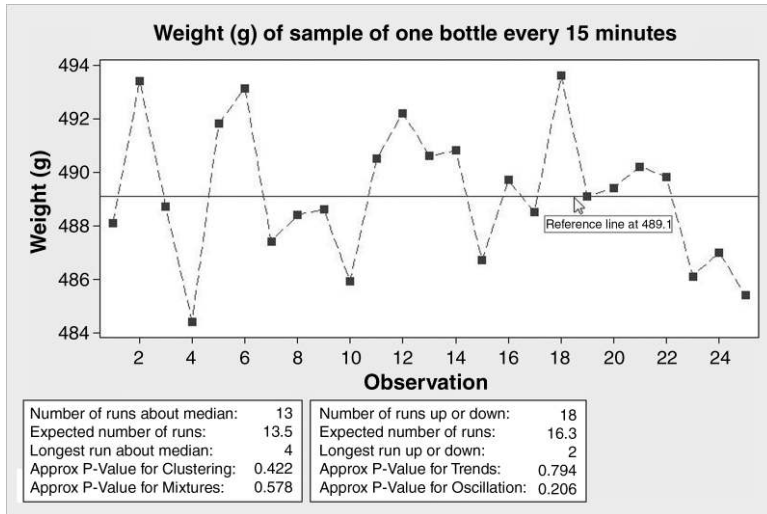


Figure 2.6 Run chart of bottle weight data.

corresponding to bottle number 19 actually on the line. Thus the median is simply the ‘middle’ of the data set. (The sample of five bottle weights 488.4, 490.8, 486.1, 489.3 and 490.5 may be ordered to yield 486.1, 488.4, **489.3**, 490.5 and 490.8 so the median is the middle value 489.3. The sample of six bottle weights 487.4, 488.1, 493.6, 492.2, 488.6 and 490.6 may be ordered to yield 487.4, 488.1, **488.6**, **490.6**, 492.2 and 493.6. In this case, where there is an even number of bottles in the sample, the median is taken as the value midway between the middle pair of values, i.e. $(488.6 + 490.6)/2 = 489.6$.)

The median of 489.1 for the sample of 25 bottles is a measure of process performance in relation to the target weight of 490 g. Faced with just this information, should the process owner take action to ‘shift’ the measure of location closer to 490? This is the sort of question on which statistics can shed light – further discussion of such questions and the tools available in Minitab to answer them appears in Chapter 7.

However, before considering the performance in relation to any target or specification limits, a much more fundamental question should be asked: is the process performing in a stable, predictable manner? The information displayed beneath the run chart is relevant. For an explanation, the Help facility provides the overview of run charts displayed in Figure 2.7. One way to access this information is to click on the **Help** button in the bottom left-hand corner of the Run Chart dialog box – see Figure 2.4. Note also the provision of [how to](#), [example](#), [data](#) and [see also](#) links to further sources of information to aid the user to learn about run charts. In addition, an explanation of the dialog box items is given and a link to information on the options available in creating a run chart in Minitab via [Options](#). Links to related topics are also provided within the text – in the case of the run chart the related topics are [subgroup](#) and [median](#).

A process performing in a stable and predictable manner is said to exhibit *common cause variation* only and to be in a state of statistical control. When a process is affected by *special cause variation*, i.e. by variation resulting from causes extraneous to the process, evidence of the presence of such variation may be provided by the tests referred to in the second paragraph of the overview. Data for a process affected only by common cause variation exhibit randomness while data for a process do not. The tests are often referred to as tests for

Run Chart
[overview](#) [how to](#) [example](#) [data](#) [see also](#)

Stat > Quality Tools > Run Chart

Use Run Chart to look for evidence of patterns in your process data, and perform two tests for non-random behavior. Run Chart plots all of the individual observations versus the subgroup number, and draws a horizontal reference line at the median. When the subgroup size is greater than one, Run Chart also plots the subgroup means or medians and connects them with a line.

The two tests for nonrandom behavior detect trends, oscillation, mixtures, and clustering in your data. Such patterns suggest that the variation observed is due to special causes – causes arising from outside the system that can be corrected. Common cause variation is variation that is inherent or a natural part of the process. A process is in control when only common causes affect the process output.

Dialog box items

Data are arranged as

Single column: Choose if data is in one column. Enter a column.

Subgroup size (use a constant or an ID column): Enter the subgroup size (for equally sized subgroups) or a column of subscripts (for unequally sized subgroups).

Subgroups across rows of: Choose if subgroups are arranged in rows across several columns. Enter the columns.

For data in subgroups You can plot either the subgroup means or medians as points on the graph. Minitab uses the points to count the number of runs in tests for randomness.

Plot subgroup means: Choose to plot the subgroup means as points on the graph.

Plot subgroup medians: Choose to plot the subgroup medians as points on the graph.

<Options>

Figure 2.7 Help on run charts.

randomness. In order to conduct these one must scrutinize the P -values in the text boxes beneath the run chart. P -values will be explained in Chapter 7, but at this stage one need only know that it is generally accepted that any P -value less than significance level or α -value of 0.05 provides evidence of the presence of special cause variation, i.e. of the presence of a factor or factors affecting process performance. For the weight data none of the P -values is less than 0.05 so it would appear that the bottle production process is behaving in a stable, predictable manner as far as the mould from which the bottles were sampled is concerned.

The run charts in Figures 2.8–2.11 display weight data for moulds where the tests do provide evidence of special cause variation. In the first scenario, displayed in Figure 2.8,

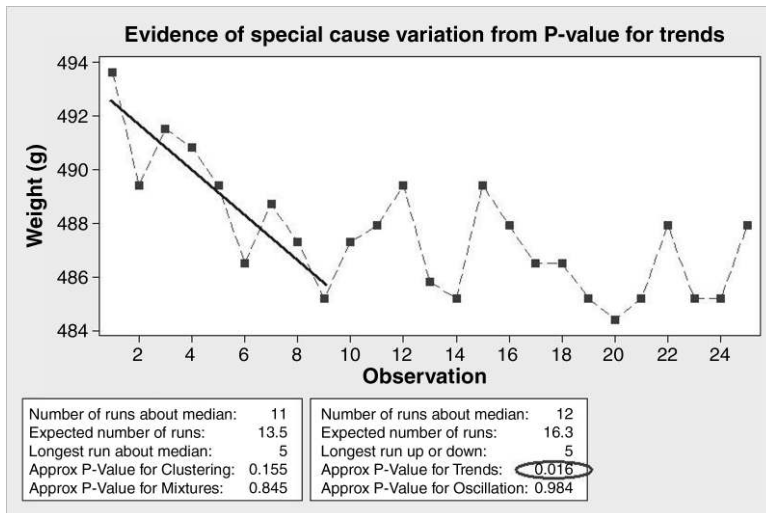


Figure 2.8 Evidence of special cause variation – scenario 1.

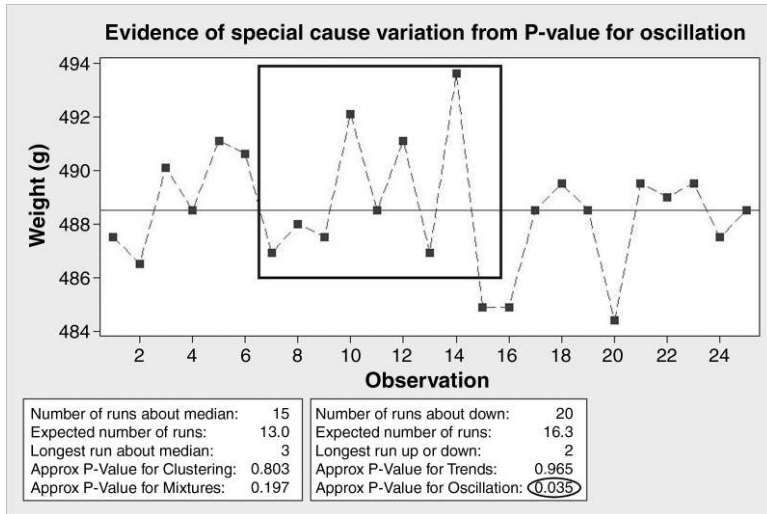


Figure 2.9 Evidence of special cause variation – scenario 2.

the *P*-value for trends is 0.016, which is less than 0.05, so there is evidence of the presence of special cause variation affecting the corresponding mould. The line segment superimposed on the display indicates an apparent initial downward trend in weight.

In the second scenario, displayed in Figure 2.9, the *P*-value for oscillation is 0.035, which is less than 0.05, so there is evidence of the presence of special cause variation affecting the corresponding mould. The rectangle superimposed on the display indicates a period during which weight oscillates rapidly.

In the third scenario, displayed in Figure 2.10, the *P*-value for mixtures is 0.012, which is less than 0.05, so there is evidence of the presence of special cause variation affecting the

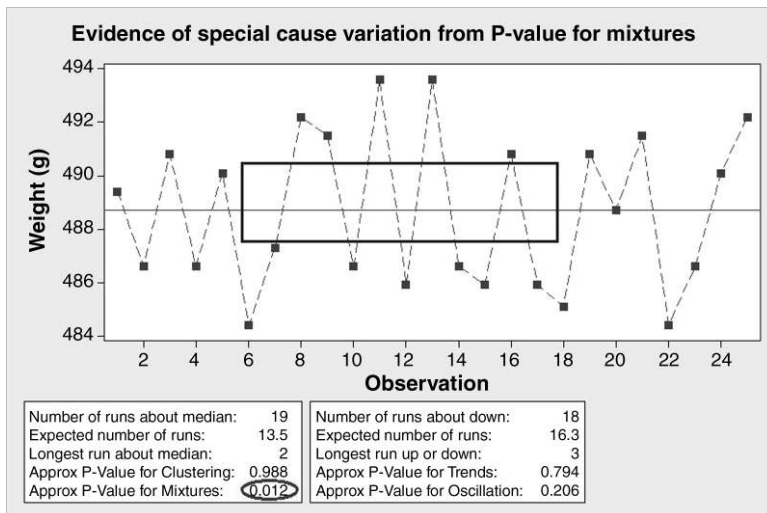


Figure 2.10 Evidence of special cause variation – scenario 3.

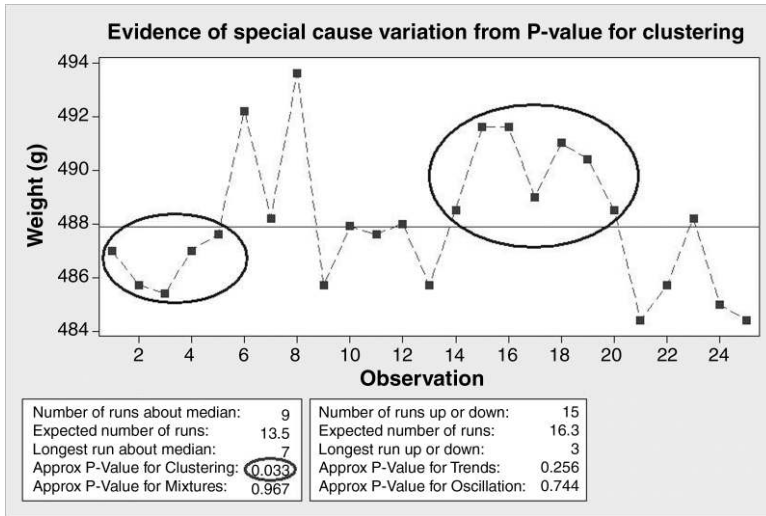


Figure 2.11 Evidence of special cause variation – scenario 4.

corresponding mould. The rectangle superimposed on the display indicates a period during which there is an absence of weight values close to the median weight represented by the reference line – typical when mixtures occur.

In the fourth scenario, displayed in Figure 2.11, the *P*-value for clustering is 0.033, which is less than 0.05, so there is evidence of the presence of special cause variation affecting the corresponding mould. Two clusters – groups of points corresponding to bottles with similar weights – are indicated in the display.

A process team should respond to evidence of special cause variation by taking steps to carry out a root cause investigation in order to determine the extraneous factor or factors affecting process performance. Once any such factor or factors have been identified, steps may be taken to eliminate them. It should be noted that a signal of evidence of the presence of special cause variation from a run chart *P*-value less than 0.05 could arise purely by chance, even when a process is operating in a stable and predictable manner.

The reader is urged to tap the huge Minitab Help resource constantly. Further details are provided in Chapter 11, and the author suggests that it will be beneficial to refer to these details in parallel with study of this and later chapters. Returning to the Minitab session currently being described, it should be noted that use of **Edit > Copy Graph** enables a copy of the run chart to be copied and pasted into a document being prepared using word-processing software. Alternatively, **File > Save Graph As...** may be used to save the run chart in a variety of formats. Minimize the run chart and note how the text **run chart of Weight (g)** has appeared in the Session Window indicating that in the Minitab session to date a run chart of the weight data has been created.

2.1.2 Minitab projects and their components

View the Project Manager either by clicking on the icon second from the right on the Project Manager toolbar displayed in Figure 2.2, using the Project Manager icon at the bottom left of the screen or by using keystrokes (Ctrl + I). The display in Figure 2.12 results.

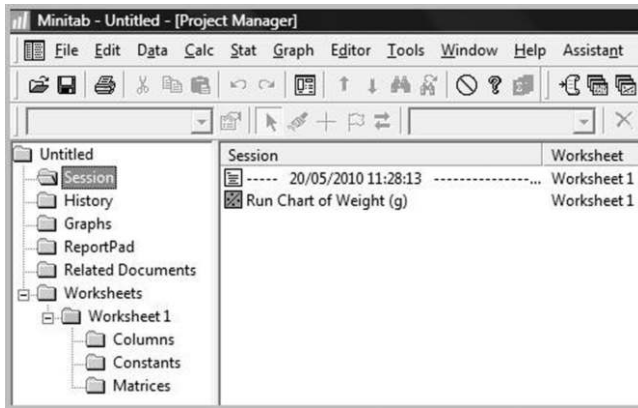


Figure 2.12 Project Manager.

In the top left-hand corner the word ‘Untitled’ indicates that the Project has not yet been named. The contents of the open Session folder indicate the date and time of the creation of Worksheet1 and the subsequent display of the data in the run chart. The run chart is in the Graphs folder. On opening the ReportPad folder, a report document may be created with appropriate text being entered as shown in Figure 2.13. Subsequently the run chart may be inserted into the ReportPad using **Edit > Paste** if the **Edit > Copy Graph** option was used

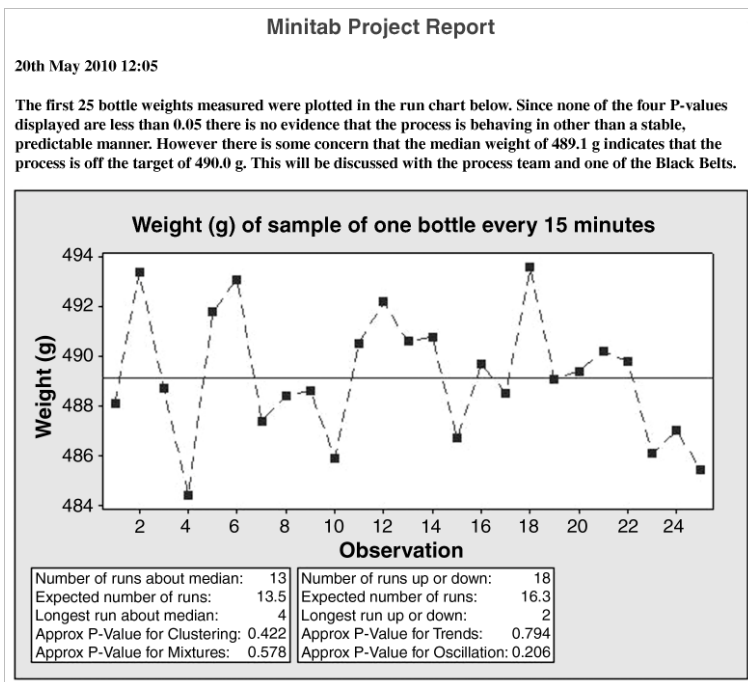


Figure 2.13 ReportPad showing text and run chart.

Table 2.2 Additional bottle weight data.

Sample	Weight	Sample	Weight	Sample	Weight
26	484.5	31	487.0	36	486.4
27	489.9	32	485.1	37	485.3
28	485.0	33	486.3	38	486.9
29	485.9	34	490.0	39	486.7
30	488.1	35	485.2	40	484.6

earlier. Alternatively, on right-clicking the active run chart a menu appears; clicking the option **Append Graph to Report** adds the chart directly to the ReportPad.

The typical final step in such a first session with data from a process is the naming and saving of the Minitab project file. To achieve this use **File > Save Project As...** to save the project with the name Weight in an appropriate folder. The project file will be created as Weight.MPJ, with the extension .MPJ indicating the file type as a Minitab project. **File > Exit** closes down Minitab – you should do just that and take a well-earned rest! To continue working with some other data in a new Minitab project, when one has finished work on a current project and saved it, use **File > New > Minitab Project**. One may use **File > New > Minitab Worksheet** to create additional worksheets within a project.

Imagine that a discussion took place with the process manager on concern about bottle weight being on target and that he consults a Six Sigma Black Belt, who does some further analysis of the available data using Minitab and reassures the process team that there is no evidence to suggest that the process is off target. As production of the batch of bottles continues, further data became available which are displayed in Table 2.2.

Launch Minitab. Use **File > Open Project** to open the project file Weight.MPJ created and saved earlier. The Toolbar at the top of the screen may be used to access components of the Project as indicated earlier (Figure 2.2). Click on the Current Data Window icon (or on the Show Worksheets Folder icon) to access the only worksheet currently in the project. Add the additional data to the first column of the worksheet. Using: **Stat > Quality Tools > Run Chart...** (with **Subgroup size: 1**), a run chart of the updated data set may now be created in order to make a further check on process performance as the production of bottles continues.

The updated chart is displayed in Figure 2.14. The *P*-value for clustering of 0.028 is less than 0.05, so therefore there is evidence of a possible special cause affecting the process. Scrutiny of the run chart reveals that the additional data points form a cluster, and scrutiny of the actual data values in Table 2.2 indicates that all but one of the additional bottles had weight less than the target value of 490 g. Thus it would appear that corrective action, to remove a special cause of variation affecting the process, could be necessary. On accessing the ReportPad via its icon one can type appropriate further comments and add the updated run chart as shown in Figure 2.14.

Bottle weight is a key process output variable in this context. People involved with the process will have the knowledge of the key process input variables that can be adjusted in order to bring weight back to the desired target of 490 g. One might state, from scrutiny of the second run chart, that the ‘drop’ in weight is obvious. There was a ‘cluster’ of 25 bottles initially with median weight of 489.1 g and a later cluster of 15 bottles with a median weight of 486.3 g. (You can readily check the second median by putting the data in Table 2.2 in order and picking out the middle value; the calculation of medians using Minitab will be covered later in the chapter.) However, the objective evaluation of evidence from data using sound statistical methods is preferable to subjective decision-making.

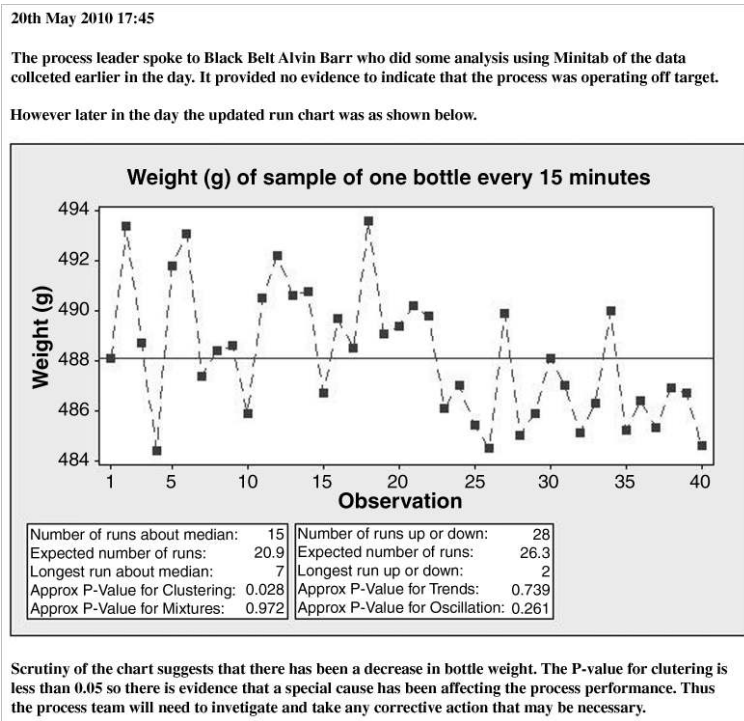


Figure 2.14 ReportPad showing run chart of the 40 bottle weights and further comments.

Another useful source of information in Minitab is StatGuide™. Right-clicking on an active run chart and then clicking on **StatGuide** opens a window with a display of **Contents** on the left of the screen and with **Index** and **Search** tabs. On the right specific information on run charts is given. Arrows enable navigation around the topics provided, example output is given, and the location of the data used to create it is indicated together with interpretation. The **More** button leads to in-depth details of how the interpretation is made.

The reader will find further details of Minitab StatGuide in Chapter 11, and the author suggests that it will be beneficial to refer to these details in parallel with study of this and later chapters.

Having completed your work with the bottle weight data discussed above, the natural thing to do would be to save the updated Minitab project file Weight.MPJ using **File > Save Project**. Were the project ‘for real’ then the worksheet could be updated as new data became available and the ReportPad could be used as a dynamic document containing informative displays and analyses of the data and a log of any changes made to the process. Worksheets may be stored independently of projects, in the first instance, using **File > Save Current Worksheet As...**, and subsequently updates may be saved using **File > Save Current Worksheet**. It is recommended that you save the worksheet containing the 40 weights using the name Weight1. The worksheet file will be created as Weight1.MTW, with the extension .MTW indicating the file type as a Minitab worksheet.

The response variable, weight, considered above is an example of a *continuous random variable* in the jargon of statistics. When a bottle is weighed on a set of analog scales one can think of the possibility of the pointer coming to rest at any point on a continuous scale of

measurement. For a second example of a run chart the number of incomplete invoices per day produced by the billing department of a company will be considered. The daily count of incomplete invoices is referred to as a *discrete random variable* in statistics. It should be noted that both weighing bottles and counting incomplete invoices are examples of measurement.

The majority of the data sets used in this book may be downloaded from the web site http://www.wiley.com/go/six_sigma in the form of Minitab worksheets or Excel workbooks. It is recommended that you download the files and store them in a directory on your computer. The data for this example, available in Invoices1.MTW, are from 'Finding assignable causes' by Bisgaard and Kulachi and are reproduced with permission from *Quality Engineering* (© 2000 American Society for Quality).

In order to create a new project for the invoice data use **File > New**, select **Minitab Project** and click **OK**. (Had you omitted to save the updated bottle weight project you would have been offered the option of doing so on clicking **OK**. It is strongly advised that you save projects as you work your way through this book as many data sets will provide opportunities for analysis using other methods in later chapters.) A new blank project file is opened. In order to save the reader the tedious task of typing in the initial invoice data displayed in Figure 2.15 the data can

↓	C1-D	C2	C3
	Date	No. Invoices	No. Incomplete
1	03/01/2000	98	20
2	04/01/2000	104	18
3	05/01/2000	97	14
4	06/01/2000	99	16
5	07/01/2000	97	13
6	10/01/2000	102	29
7	11/01/2000	104	21
8	12/01/2000	101	14
9	13/01/2000	55	6
10	14/01/2000	48	6
11	17/01/2000	50	7
12	18/01/2000	53	7
13	19/01/2000	56	9
14	20/01/2000	49	5
15	21/01/2000	56	8
16	24/01/2000	53	9
17	25/01/2000	52	9
18	26/01/2000	51	10
19	27/01/2000	52	9
20	28/01/2000	47	10

Figure 2.15 Initial invoice data.

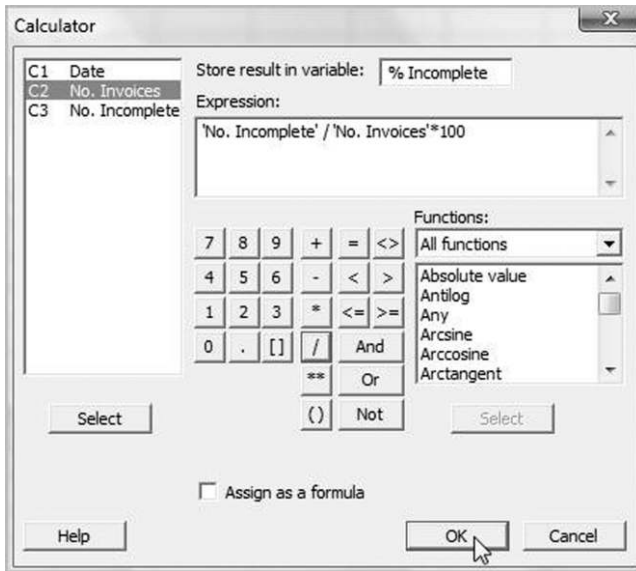


Figure 2.16 Calculator dialog box.

be entered into the project using **File > Open Worksheet**. Select the file *Invoices1.MTW* from the directory in which you have stored the downloaded data sets and click on **Open**. (You may be asked for confirmation that you wish to add a copy of the content of the worksheet to the current project, in which case it is necessary to click **OK**.)

There are three columns in the worksheet displayed in Figure 2.15. The first is labelled C1-D, indicating that it contains date data. The columns labelled C2 and C3, with no extensions, hold numerical data – the daily number of invoices processed and the daily number of invoices found to be incomplete. A run chart of the number of incomplete invoices per day could be misleading since the number of invoices processed daily varies. Thus there is a need to calculate the proportion of incomplete invoices per day. Using **Calc > Calculator...** gives access to the dialog box displayed in Figure 2.16 for performing calculations in Minitab.

In the **Store result in variable:** window an appropriate name for the new column of percentages to be calculated is entered; %Incomplete was used here. In the window labelled **Expression:** the formula may be created by highlighting the names of columns, using the **Select** button and the calculator keypad. Clicking **OK** implements the calculation. (Note that a menu of functions is available for use in more advanced calculations. Note too that if one checks the **Assign as a formula** box then whenever additional data are entered in the second and third columns the percentage incomplete will be calculated automatically. Columns that have been assigned a formula are indicated by a green cross at their heads.) The run chart of %Incomplete displayed in Figure 2.17 was obtained.

Moving the mouse pointer to the horizontal reference line, representing the median percentage incomplete on the chart, triggers display of a text box giving the median as 16.1% (to one decimal place). Thus the current performance of the invoicing process is such that approximately one in every six invoices is incomplete. The *P*-value for

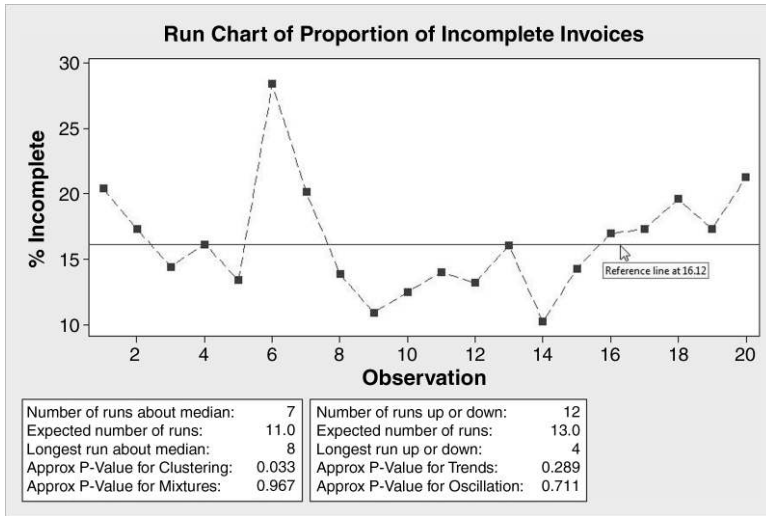


Figure 2.17 Run chart of daily percentage of incomplete invoices.

clustering is less than 0.05, thus providing evidence of the possible influence of a special cause on the process. The percentage for the sixth day appears to be considerably higher than all the other percentages – in fact, a new inexperienced employee processed many of the invoices during that day.

A median of 16.1% was unacceptably high, so a Six Sigma process improvement project was undertaken on the invoicing process, with some process changes being made almost immediately. The run chart in Figure 2.18 shows the data for the year 2000 up to

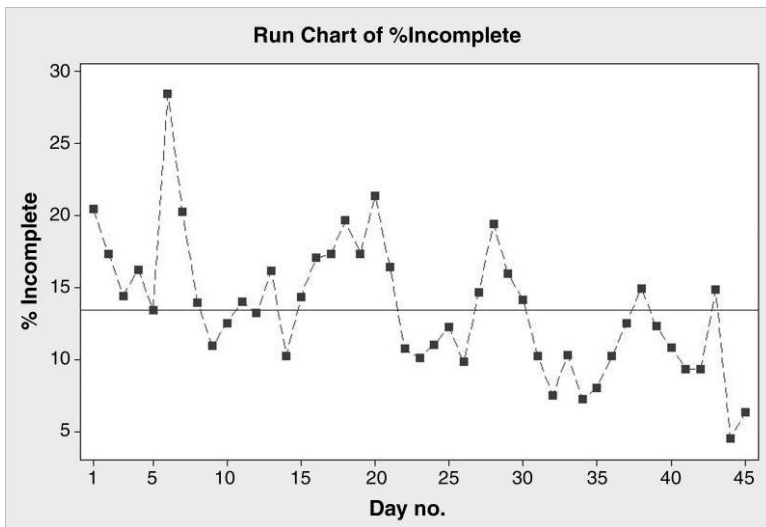


Figure 2.18 Updated run chart of daily percentage of incomplete invoices.

March 3rd. As an exercise the reader is invited to access, in a new project, the updated data stored in the worksheet Invoices2.MTW and re-create this run chart. In doing the calculation of percentage incomplete on this occasion, start to create the **Expression:** required by selecting the Round function under **Functions:**. Choose **All functions** from the menu (if not already on view), scroll down the list, highlight **Round** and click **Select**. The formula that appears in the **Expression:** window is ROUND(number,decimals). By highlighting the word ‘number’ in the expression, use may be made of highlighting and selection of column names and of the keypad to create the desired formula. The phrase ‘decimals’ may be highlighted and the digit 1 typed to indicate that rounding of the calculated percentages to one decimal place is required. The final version of the formula in the Expression window is ROUND(‘No. Incomplete’/‘No. Invoices’*100,1). On clicking **OK** the proportions of incomplete invoices as percentages will be calculated and rounded to one decimal place.

The updated run chart in Figure 2.18 was simplified by deleting the two text boxes containing the *P*-values, by clicking on each and then pressing the delete key, and editing the default label for the horizontal axis from the default of **Observation** to **Day no.** in order to make the display less daunting for use in a presentation. Double-clicking an axis label yields an Edit Dialog Label box in which any desired label may be entered in the **Text:** window.

Scrutiny of the updated run chart suggests that the process changes have been effective in lowering the percentage of incomplete invoices. Two of the *P*-values are less than 0.05. This provides formal evidence of a process change having taken place. Alternative data displays, using methods discussed later in the chapter, may be used to highlight the apparent effectiveness of the process changes. The median for the period from 31 January 2000 onwards was 10.7% incomplete invoices per day, compared with 16.1% incomplete invoices per day for the earlier period.

2.2 Display and summary of univariate data

2.2.1 Histogram and distribution

Consider the bottle weight data displayed in the run chart in Figure 2.6. Here we recorded a single variable for each bottle so we refer to *univariate* data. (Had we measured weight and height we would have had *bivariate* data, had we measured weight, height, bore, out of vertical etc. then we would have been dealing with *multivariate* data.) The process appears to have been behaving in a stable, predictable manner during the period in which the data were collected. When a process exhibits this sort of behaviour and the measured response is a continuous variable, such as weight, then display of the data in the form of a histogram is legitimate and can be very informative.

Once a bottle has been weighed, imagine that it is put in one of the series of bins depicted in Figure 2.19 according to its measured weight. The lightest bottle recorded weighed 484.4 g and would be placed in the bin labelled 483.5, 484.5. The second lightest bottle weighed 485.4 g and would be placed in the bin labelled 484.5, 485.5. The next two lightest bottles weighed 485.9 g and 486.1 g and would be placed in the bin labelled 485.5, 486.5. The heaviest bottle weighed 494.5 g and would be placed in the bin labelled 493.5, 494.5. (The convention adopted in Minitab is that a bottle weighing 490.5 g is placed in the bin labelled 490.5, 491.5 unless it

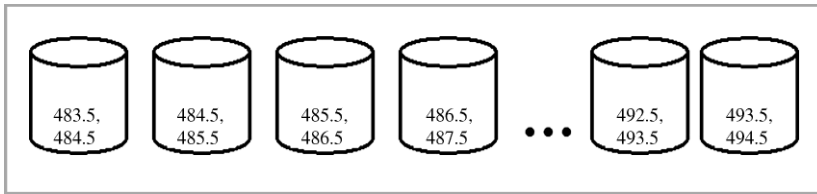


Figure 2.19 Bin concept.

was the heaviest bottle, in which case it would be placed in the bin labelled 489.5, 490.5.) For the complete sample of 25 bottles the number of bottles or observations in each bin is known as its *frequency*. The ranges 483.5–484.5, 484.5–485.5, 485.5–486.5 etc. are referred to as *intervals*. A chart with weight on the horizontal axis and frequency represented on the vertical axis by contiguous bars is a *histogram*.

In order to work through the creation of the histogram with Minitab you require the weights in Table 2.1 in a single column that may be named Weight (g). They are provided in worksheet Weight1A.MTW. To create the histogram with Minitab, use **Graph > Histogram...** The initial part of the dialog is displayed in Figure 2.20. Accept the default option of **Simple** and click **OK** to access the subdialog box displayed in Figure 2.21.

In the **Graph variables:** window, select the variable to be displayed in the histogram, Weight (g) in this case, and click **OK**. The histogram displayed in Figure 2.22 indicates the distribution of the bottle weights, and three aspects of a distribution may be assessed using a

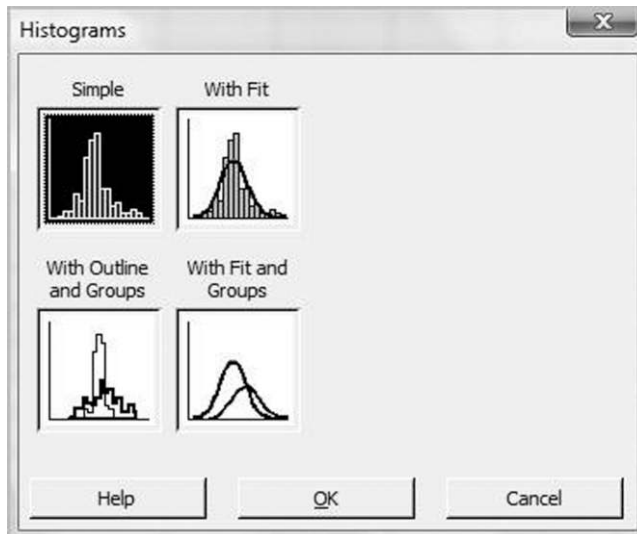


Figure 2.20 Initial part of dialog for creating a histogram.

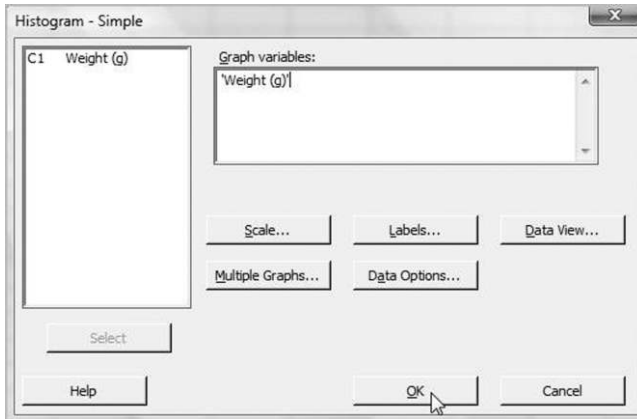


Figure 2.21 Histogram subdialog box.

histogram – *shape, location* and *variability*. Note that moving the mouse pointer to a bar of the histogram leads to the bin interval and frequency for that bar being displayed in a text box. In Figure 2.22 the mouse pointer is on the bar corresponding to the bin interval 488.5, 489.5. The frequency for this bin was 5, indicating that five of the 25 bottles in the sample had weight in the interval $488.5 \text{ g} \leq \text{weight} < 489.5$.

Before reading any further the reader is invited to create a histogram of the bottle weight data stored in the supplied worksheet Weight2.MTW. The histogram will be referred to below and is shown in Figure 2.23.

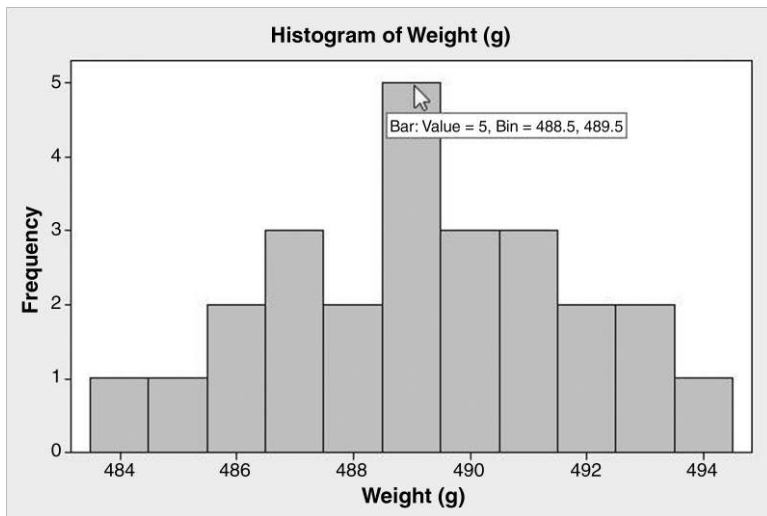


Figure 2.22 Histogram of bottle weight.

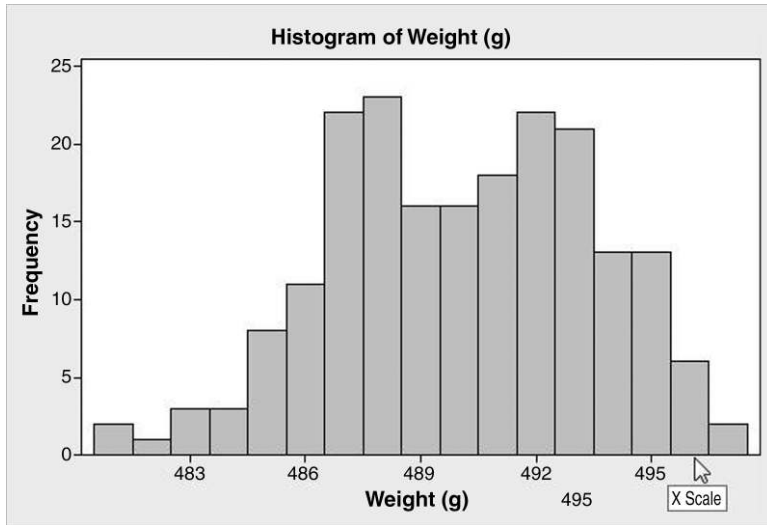


Figure 2.23 Bimodal histogram.

2.2.2 Shape of a distribution

In discussing the shape of a distribution one should consider whether or not the distribution is symmetrical, whether it is skewed to the right (showing upwards straggle) or skewed to the left (showing downwards straggle) or whether there are other features providing insights with potential to lead to quality improvement.

The histogram of weight for a sample of bottles in Figure 2.23 is bimodal – it has two major peaks. This may indicate that the sample includes bottles formed by two moulds, that the process has been run in different ways by the two shift teams responsible for its operation, etc. The third bin has midpoint 483 and the corresponding bin range or interval is 482.5–483.5, and Minitab refers to the values defining intervals as *cutpoint* positions.

In order to change the bins used in the construction of the histogram in Figure 2.23, with the graph active, move the mouse pointer to and, if necessary, along the horizontal X axis and double-click when the text box displaying the text **X Scale** appears. Select the **Binning** tab, **Interval Type Cutpoint** and **Interval Definition by Midpoint/Cutpoint positions**. Editing the list of cutpoints to become 480 485 490 495 500, as displayed in Figure 2.24, and then clicking **OK**, **OK** yields the histogram in Figure 2.25.

Note that the bimodal nature of the distribution of bottle weights is no longer evident. Thus potentially important information in a data set may be masked by inappropriate choice of binning intervals. Further reference to distribution shape will be made later in the chapter.

2.2.3 Location

In discussing location the question being addressed with regard to process performance is ‘where are we at?’ (The author prefers to use ‘location’ rather than alternative term ‘central tendency’.) Location gives an indication of what is typical in terms of process performance. Suppose that the weight data displayed in Figure 2.23 were actually for 100 bottles produced on each of two different moulding machines, A and B; that it is known which bottles were made

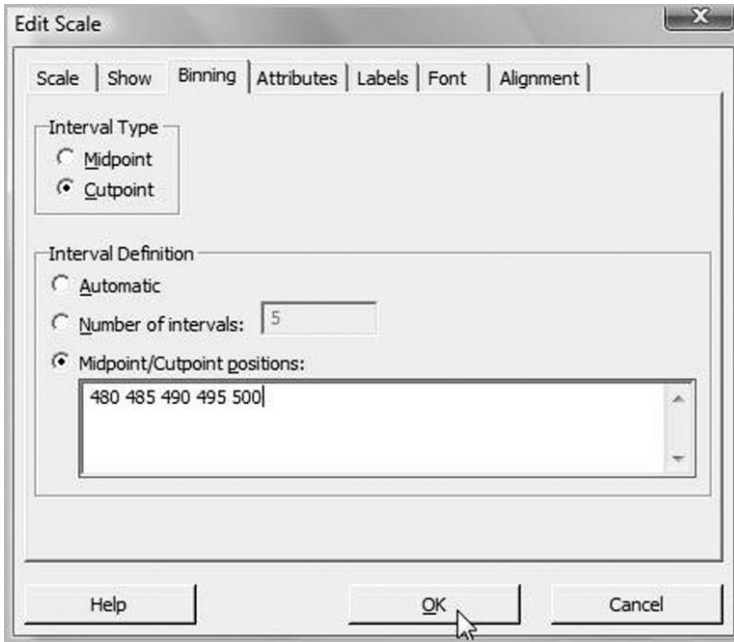


Figure 2.24 Changes to the bin cutpoint positions.

on each machine and that bottle weight for both was stable and predictable during the period when the samples were taken. Part of the supplied worksheet, Weight3.MTW, containing the data is displayed in Figure 2.26. It shows the final four bottle weights for machine A and the first four bottle weights for machine B.

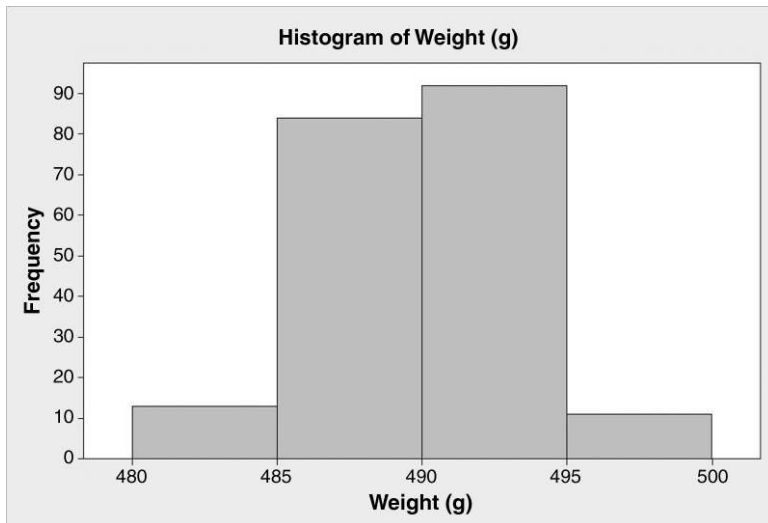


Figure 2.25 Alternative histogram.

↓	C1-T	C2
	Machine	Weight (g)
97	A	488.5
98	A	487.2
99	A	486.5
100	A	485.5
101	B	492.0
102	B	491.6
103	B	494.5
104	B	493.8

Figure 2.26 Segment of bottle weight data for two machines.

Column C1 contains text values A and B indicating which of the two machines produced the bottle with weight recorded in column C2. Note the designation of the first column as C1-T, indicating that it stores text values. In order to create a histogram for each machine use **Graph > Histogram...** with the **Simple** option and select the variable to be graphed, **Weight (g)**. Click on **Multiple Graphs...** and **By Variables** and select Machine in the window labelled **By variables with groups in separate panels:** as displayed in Figure 2.27. Finally click **OK, OK**.

The two histograms are shown in Figure 2.28. The histogram for machine A is in the left-hand panel and that for machine B is in the right-hand panel. With the graph active, the **Edit Scale** menu was accessed by moving the mouse pointer to the horizontal X axis and double-clicking when the text box displaying the text **X Scale** appeared. The entries in the window labelled **Positions of ticks:** were changed to 480, 485, 490 and 495. The triangular markers were superimposed using the polygon tool from the Graph Annotation Tools toolbar, but the detail need not concern us here. These marks indicate the mean weight for each machine. The mean will be defined later in this chapter. The markers indicate the horizontal locations of the centroids of the histograms. Cut-outs of the histograms would balance on the knife-edges represented by the upper vertices of these triangles. In terms of the target weight of 490 g for the bottles, it is clear from the data display that both machines are operating off target.

The difference in location for the two machines and in their performance, relative to the target bottle weight of 490 g, may be highlighted as shown in Figure 2.29 with the histograms aligned vertically. Details of how to do this will be the subject of an exercise at the end of this chapter.

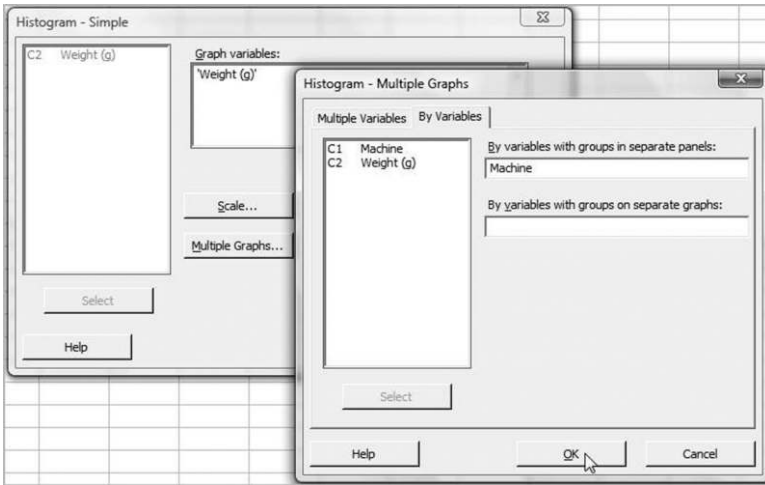


Figure 2.27 Segment of bottle weight data for two machines.

In addition to visual assessment of location from display of the data it is possible to measure location by calculation of descriptive or summary statistics. The median is a widely used measure of location and was referred to in the previous section in relation to run charts. The mean is a second widely used measure of location and is obtained by calculating the sum of data values in a sample and dividing by the sample size, i.e. by the number of data values. In common parlance many refer to the mean as the average. Calculation of the mean with associated statistical notation is given in Box 2.1.

The means of bottle weight for machines A and B are 487.24 g and 492.75 g, respectively. The triangular markers in Figure 2.29 are placed at these values on the appropriate scales. The

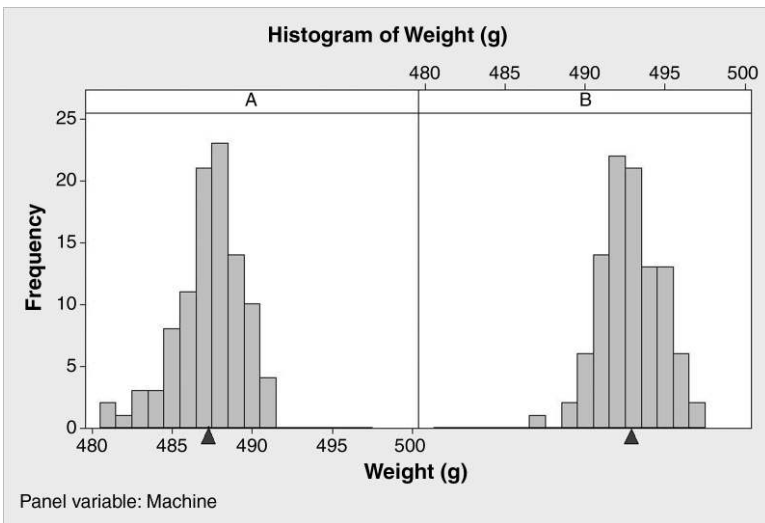


Figure 2.28 Histograms for two machines.

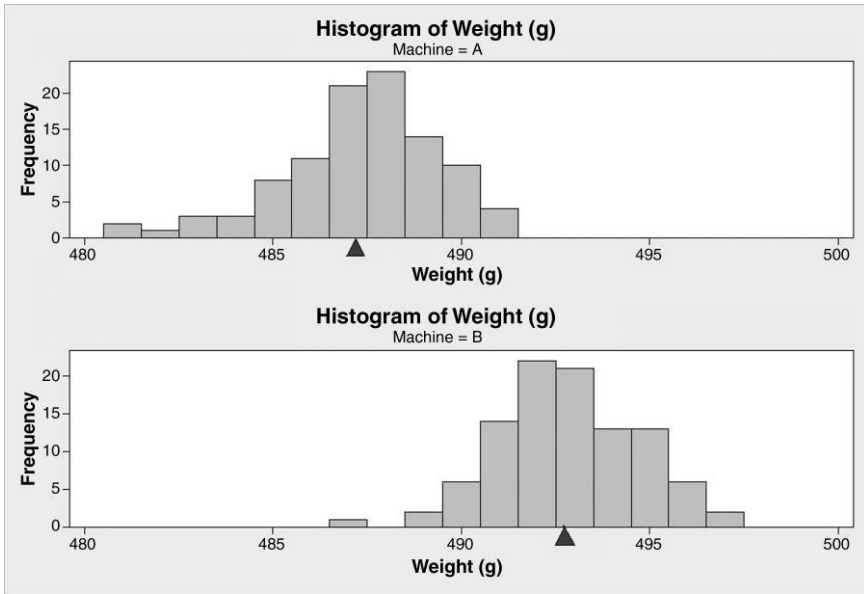


Figure 2.29 Histograms for two machines.

Consider a sample of four bottles with weights (g) 490.3, 489.9, 490.6 and 490.0. The sum of the four data values is 1960.8 and division by the sample size, 4, gives the mean weight 490.2. The mathematical shorthand for this calculation is as follows:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

where \bar{x} denotes the mean of x and Σ is the upper case Greek letter sigma denoting 'sum of'. For our sample of 4,

$$\bar{x} = \frac{\sum_{i=1}^4 x_i}{4}$$

(the $i = 1$ and the 4 indicate that the sum of the measurements, x_i , labelled 1 to 4 inclusive, is to be computed)

$$\begin{aligned} \bar{x} &= \frac{x_1 + x_2 + x_3 + x_4}{4} \\ &= \frac{490.3 + 489.9 + 490.6 + 490.0}{4} \\ &= \frac{1960.8}{4} = 490.2. \end{aligned}$$

Box 2.1 Calculation of a mean.

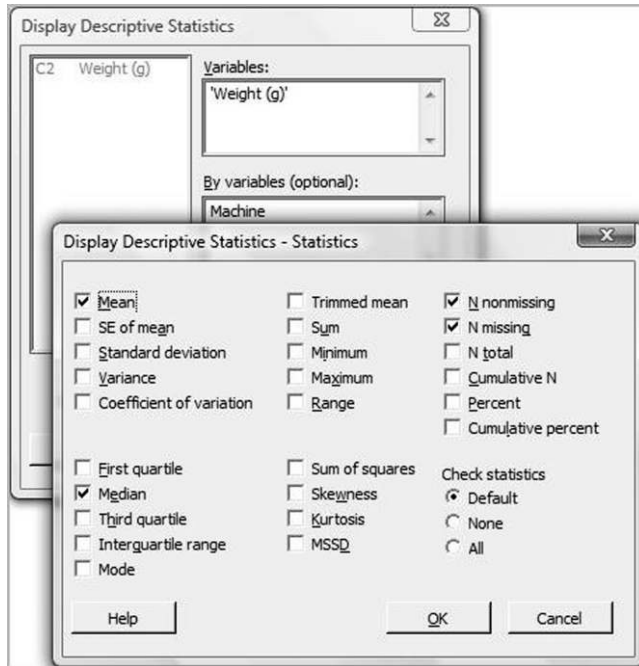


Figure 2.30 Obtaining Descriptive Statistics.

median bottle weights for machines A and B are 487.5 g and 492.7 g, respectively. With an even sample size of 100, the median is the mean of middle two weights when the data have been ordered. Note that the means and medians for each machine are very similar: 487.24 and 487.5 for A and 492.75 and 492.7 for B, respectively. This is typical when distributions (and associated histograms) are fairly symmetrical. In such cases it would not really matter which measure of location is used to summarize the data.

In order to obtain measures of location in Minitab use can be made of the **Stat** menu via the sequence **Stat > Basic Statistics > Display Descriptive Statistics...** The icon to the left of the **Display Descriptive Statistics...** text shows the two symbols \bar{x} and s , representing mean and standard deviation, respectively. The standard deviation is a widely used measure of variability which will be introduced later in this chapter. Weight (g) is entered under **Variables:** and Machine entered in **By variables:**.

In order to obtain the mean and median for the two machines, select Weight (g) in the **Variables:** window and use the **Statistics** button to edit the list of available **Statistics** to the ones shown in Figure 2.30, i.e. **Mean, Median, N nonmissing** and **N missing**. On implementation of the procedure the output in Panel 2.1 appears in the Session Window.

Descriptive Statistics: Weight (g)					
Variable	Machine	N	N*	Mean	Median
Weight (g)	A	100	0	487.24	487.50
	B	100	0	492.75	492.70

Panel 2.1 Session window output from Descriptive Statistics.

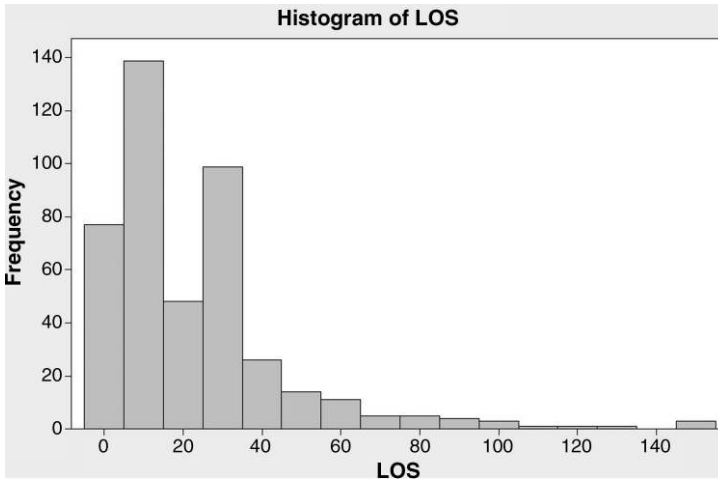


Figure 2.31 Histogram of length of stay in hospital.

The N column indicates that the data set includes values of weight for 100 bottles from each of the machines, A and B. The N* column indicates that no missing values were recorded – missing values are recorded in Minitab as asterisks. The final two columns give the means and medians. Note that, with the mouse pointer located in the Descriptive Statistics section in the Session window, a right click displays a pop-up menu through which access to StatGuide information on Descriptive Statistics may be obtained.

Consider now data on length of stay (days) in hospital (LOS) for stroke patients admitted to a major hospital during a year. The data are available in LOS.MTW. A histogram of the data is shown in Figure 2.31. Such data could be highly relevant during the measure phase of a Six Sigma project aimed at improving stroke care in the hospital.

The histogram is far from symmetrical. With the long tail to the right it exhibits what is known as *positive skewness* or *upward straggle*. (A histogram that had the shape of the mirror image of the one in Figure 2.31 in the vertical axis would exhibit *negative skewness* or *downward straggle*.) Scrutiny of the bars indicates that the bins used are $(-5, 5)$ $(5, 15)$ $(15, 25)$ etc., with midpoints 0, 10, 20 etc. Of course LOS cannot be a negative number, so a more logical set of bins would be $(0, 10)$ $(10, 20)$ $(20, 30)$ etc. In order to modify the histogram select the **X Scale** as indicated in the previous section, under **Binning** select **Interval Type** as **Cutpoint**, select **Interval Definition** as **Midpoint/Cutpoint positions**: and complete the dialog box as shown in Figure 2.32. This gives the histogram in Figure 2.33.

The default descriptive statistics provided by Minitab for length of stay are shown in Panel 2.2. (SE Mean denotes the standard error of the mean, and Q1 and Q3 denote the first and third quartiles respectively. These statistics are explained later in the book.) As is typical with data exhibiting positive skewness, the median is less than the mean. As the term ‘average’ is used by some to mean measure of location it is important, in the case of skewness, to ascertain which measure of location is being quoted. In this case the median length of stay is approximately one week less than the mean.

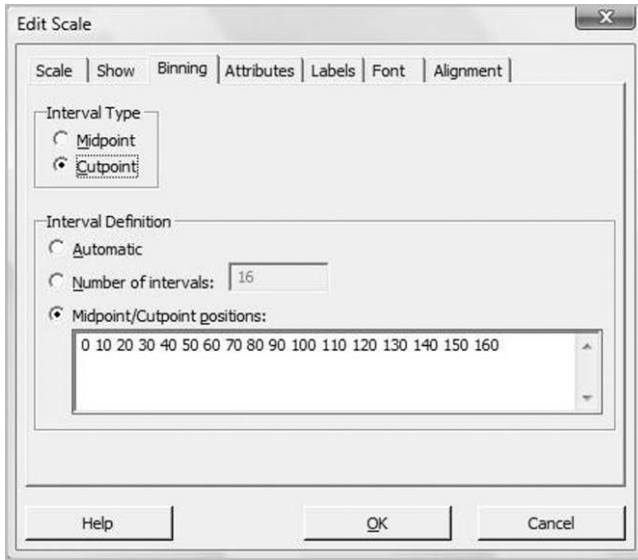


Figure 2.32 Specifying bins for a histogram using cut points.

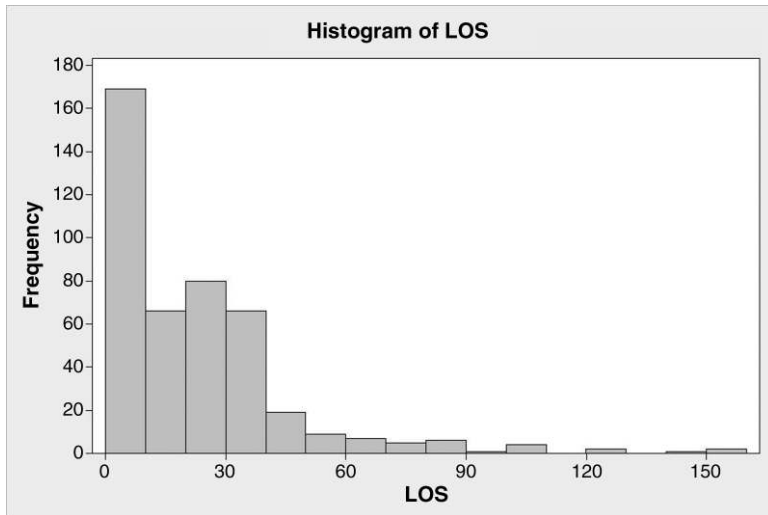


Figure 2.33 Modified histogram of length of stay in hospital.

Descriptive Statistics: LOS										
Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
LOS	437	0	22.39	1.12	23.35	1.00	6.00	15.00	31.00	154.00

Panel 2.2 Descriptive Statistics for length of stay.

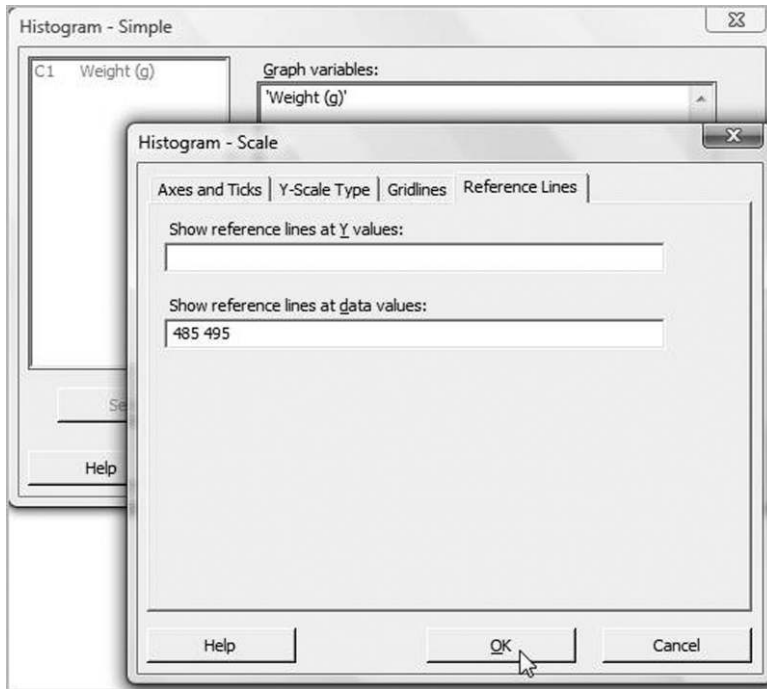


Figure 2.34 Specifying reference line positions for the data scale.

One final facility covered in this section is that of being able to add reference lines corresponding to values on the horizontal scale. This is useful for giving a visual impression of how well a process is performing in terms of customer requirements. For example, suppose that a customer of the bottle manufacturer specifies that bottle weight should lie between 485 and 495 g. Thus the customer is specifying a lower specification limit (LSL) of 485 g and an upper specification limit (USL) of 495 g. For the first sample of bottle weight data (Weight 1A.MTW) presented in this chapter, having selected Weight (g) as the variable to be graphed in the form of a histogram, click on the **Scale...** button, then on the **Reference Lines** tab and complete the dialog as shown in Figure 2.34. Clicking **OK**, **OK** twice yields the histogram with reference lines indicating the specification limits shown in Figure 2.35.

The display indicates that not all bottles met the customer specification limits – there is some ‘fall-out’ below the lower limit. In Chapter 6 indices for the assessment of how capable a process is of meeting customer specifications will be introduced.

2.2.4 Variability

In discussing variability, or spread, one is addressing the question of how much variation there is in process performance. In order to introduce measures of variability, consider two samples of five bottles from two moulding machines, P and Q. Recall that in order to create a new Minitab project, when one has finished work on a current project and saved it, one may use **File > New > Minitab Project**, and that one may use **File > New > Minitab Worksheet** to create additional worksheets within a project. Set up the data as shown in Figure 2.36. On

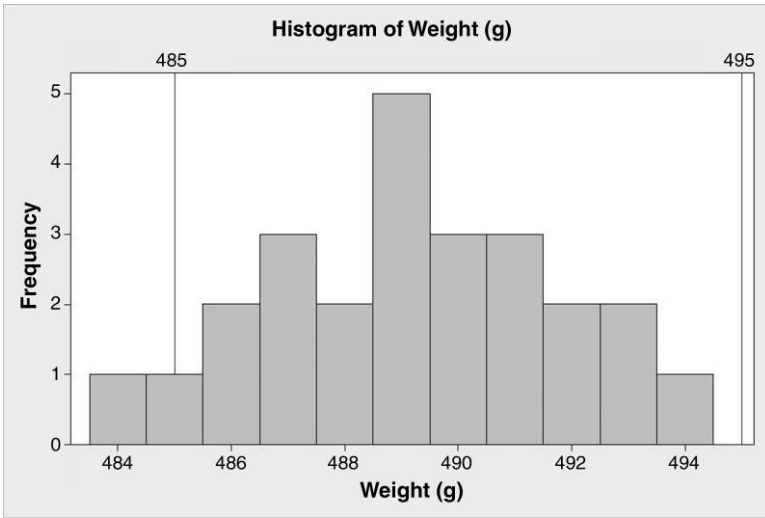


Figure 2.35 Bottle weight histogram with specification limits.

↓	C1-T	C2
	Machine	Weight (g)
1	P	488.3
2	P	491.9
3	P	489.6
4	P	487.7
5	P	492.5
6	Q	490.1
7	Q	490.2
8	Q	488.8
9	Q	491.6
10	Q	489.3

Figure 2.36 Bottle weight data from two machines.

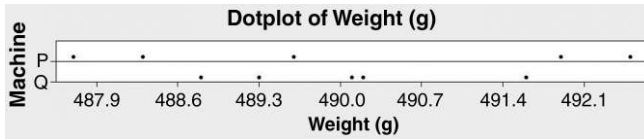


Figure 2.37 Dot plots of bottle weight by machine.

typing the letter P in the first cell of column 1 note that the column changes from C1 to C1-T, indicating that the column contains text as opposed to numeric data.

The dot plot is a useful alternative form of data display to the histogram, especially for small samples. Use **Graph > Dotplot...** to access the appropriate dialog box and select the **With Groups** option for **One Y**. Weight (g) is selected in **Graph variables:** and Machine in **Categorical variables for grouping:**. This yields the display in Figure 2.37.

Both samples have mean 490.0, but the samples differ in that the weights for machine P are more widely dispersed about the mean than are the weights for machine Q. There is greater variability, or spread, for weight in the case of machine P than in the case of machine Q. The reader is invited to verify that the default set of descriptive statistics for the two machines displayed in Panel 2.3 is obtained using **Stat > Basic Statistics > Display Descriptive Statistics...**

One measure of variability is the *range*, i.e. the difference between the minimum value in the sample and the maximum value in the sample. Using the minimum and maximum values in Panel 2.3 gives the range for machine P to be 4.8 while that for machine Q is 2.8. The greater range for machine P indicates the greater variability in the weight of bottles produced on it than the variability in the weight of bottles produced on machine Q. The range has applications in control charts for measurement data. However, one criticism of the range as a measure of variability is that it only uses two measurements from all the measurements in the sample.

The standard deviations (StDev) given in Panel 2.3 for the two machines are 2.13 and 1.07 respectively. The greater standard deviation for machine P indicates the greater variability of bottle weight for it compared with the variability of bottle weight for machine Q. Detailed explanation of the calculation of standard deviation for machine P is given in Table 2.3.

Descriptive Statistics: Weight (g)									
Variable	Machine	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median
Weight (g)	P	5	0	490.00	0.954	2.13	487.70	488.00	489.60
	Q	5	0	490.00	0.476	1.07	488.80	489.05	490.10
Variable	Machine	Q3	Maximum						
Weight (g)	P	492.20	492.50						
	Q	490.90	491.60						

Panel 2.3 Descriptive Statistics for machines P and Q.

Table 2.3 Calculation of measures of variability for weight (Machine P).

Bottle number	Weight x_i	Deviation $x_i - \bar{x}$	Absolute deviation	Squared deviation
1	488.3	-1.7	1.7	2.89
2	491.9	1.9	1.9	3.61
3	489.6	-0.4	0.4	0.16
4	487.7	-2.3	2.3	5.29
5	492.5	2.5	2.5	6.25
Total	2450.0	0	8.80	18.20
Mean	490.0	0	1.76	3.64

It is widespread practice to use the symbol x_i for a typical data value so that, for example, x_3 represents the weight of the third bottle in the sample, namely 489.6 g. The mean weight, \bar{x} , is 490.0 so the deviation of the third bottle weight from the mean is $x_3 - \bar{x} = 489.6 - 490.0 = -0.4$. This indicates that the third bottle had a weight 0.4 g below the mean. Similarly, for example, the fifth bottle had weight 2.5 g above the mean. The deviations from the mean always sum to zero.

The absolute deviation ignores the sign of the deviation and simply indicates by how much each measurement deviates from the mean. The mean absolute deviation (MAD) is 1.76 g for machine P. The reader is invited to verify that the corresponding value for machine Q is 0.76. The greater mean absolute deviation for machine P indicates the greater variability for it compared with that for machine Q.

Although MAD is a perfectly viable measure of variability it has disadvantages from a mathematical point of view. An alternative approach to taking the absolute values of deviations is to square the deviations and to take the mean of the squared deviations as a measure of variability. The mean squared deviation (MSD) for machine P is 3.64 while that for machine Q is 0.91. Once again the greater mean squared deviation for machine P indicates the greater variability for it compared with that for machine Q. However, there are two disadvantages with MSD. First, since the deviations are in units of grams (g) the squared deviations are in units of grams squared (g^2). Second, samples are generally taken from populations in order to estimate characteristics of the populations, but statistical theory shows that MSD from sample data underestimates MSD for the population sampled. For example, in the case of a production run of bottles the population of interest would be all bottles produced during that particular run.

An important measure of variability is sample *variance*, which is calculated as the sum of squared deviations divided by the number which is one less than the sample size. Thus for machine P the variance is given by $18.20/4 = 4.55$. Finally, in order to get back to the original units, the sample *standard deviation* is obtained by taking the square root of the variance. This yields a standard deviation of 2.13 g for machine P as displayed in the Minitab descriptive statistics in Panel 2.3. The reader is invited to verify that the standard deviation for machine Q is 1.07 g. The main point is that variance and standard deviation are very important measures of variability – the technical details of the underlying calculations are not important.

Table 2.4 Ratios of measures of variability.

Measure of variability	Machine P	Machine Q	Ratio
Range	4.8	2.8	1.7
Mean absolute deviation	1.76	0.76	2.3
Standard deviation	2.13	1.07	2.0

Table 2.4 gives the ratios of the measures of variability that have units of weight for the two machines. In broad terms all three measures indicate that the variability or spread of the weights for machine P is approximately twice that for machine Q. Small artificial samples were used here for illustrative purposes. One would be very wary of claiming that there is a real difference in variability in the weights of bottles produced on the two machines on the basis of such small samples. Readers who wish may examine the mathematics of the standard deviation in Box 2.2.

Consider again the sample of four bottles, referred to in Box 2.1, with weights (g) 490.3, 489.9, 490.6 and 490.0. The mean \bar{x} is 490.2. The mathematical shorthand for the calculation of the sample variance, s^2 , and standard deviation s , is:

$$\begin{aligned}
 s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \\
 &= \frac{\sum_{i=1}^4 (x_i - \bar{x})^2}{4 - 1} \\
 &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2}{3} \\
 &= \frac{(490.3 - 490.2)^2 + (489.9 - 490.2)^2 + (490.6 - 490.2)^2 + (490.0 - 490.2)^2}{3} \\
 &= \frac{0.1^2 + (-0.3)^2 + 0.4^2 + (-0.2)^2}{3} \\
 &= \frac{0.01 + 0.09 + 0.16 + 0.04}{3} \\
 &= \frac{0.3}{3} \\
 &= 0.1 \\
 s &= \sqrt{0.1} = 0.316.
 \end{aligned}$$

Box 2.2 Calculation of a standard deviation.

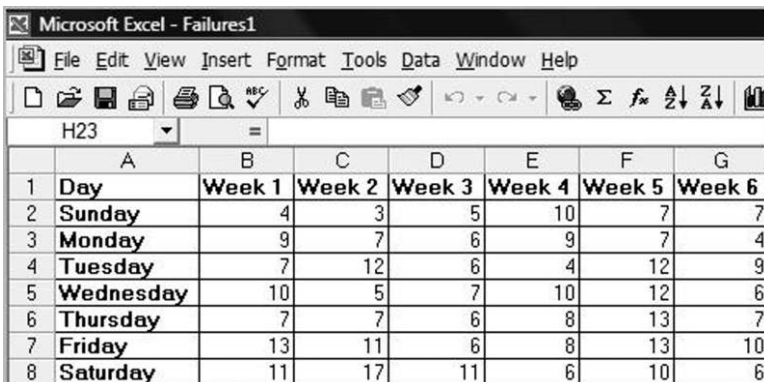
The reader is invited to set the data from Box 2.2 up in Minitab and to check the values obtained for the mean and standard deviation. Population mean and population standard deviation are widely denoted by the Greek symbols μ and σ , respectively. The sample mean and the sample standard deviation are generally denoted by \bar{x} and s , respectively. The sample values \bar{x} and s may be considered to provide estimators of the corresponding population parameters μ and σ . In statistics, a fairly general convention is to denote population values (parameters) by Greek letters and sample values (statistics) by Latin letters.

2.3 Data input, output, manipulation and management

2.3.1 Data input and output

Earlier in this chapter we saw the three types of data used in Minitab – *numeric*, *text* and *date/time*. Data may be stored in the form of columns, constants or matrices. The latter two forms will be introduced later in the book. The key scenario is one of variables in columns and cases in rows stored in a worksheet. The fundamental method of data entry is via the keyboard directly into the Data window. Once data have been entered in this way they may be stored using **File > Save Current Worksheet As...** in the case of a new worksheet or **File > Save Current Worksheet** in the case where further data has been added or changes made to an existing worksheet. Data may also be accessed via Minitab worksheet and project files created previously or from files of other types. One may use **File > Print Worksheet...** to obtain output, on paper, of the data in an active worksheet.

In order to introduce aspects of both data input and manipulation consider the data displayed in Figure 2.38 that are available as the Excel workbook Failures1.xls. It gives, for a period of 6 weeks, the number of units per day that fail to pass final inspection in a manufacturing operation that operates 24 hours a day, 7 days a week. The data set used in this example is relatively small, as are those in examples to follow, and some tasks carried out using Minitab could be done more quickly by simply retyping the data in a new worksheet! However the aim is to use small data sets to illustrate useful facilities in Minitab for the input and manipulation of data.



	A	B	C	D	E	F	G
1	Day	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
2	Sunday	4	3	5	10	7	7
3	Monday	9	7	6	9	7	4
4	Tuesday	7	12	6	4	12	9
5	Wednesday	10	5	7	10	12	6
6	Thursday	7	7	6	8	13	7
7	Friday	13	11	6	8	13	10
8	Saturday	11	17	11	6	10	6

Figure 2.38 Spreadsheet data.

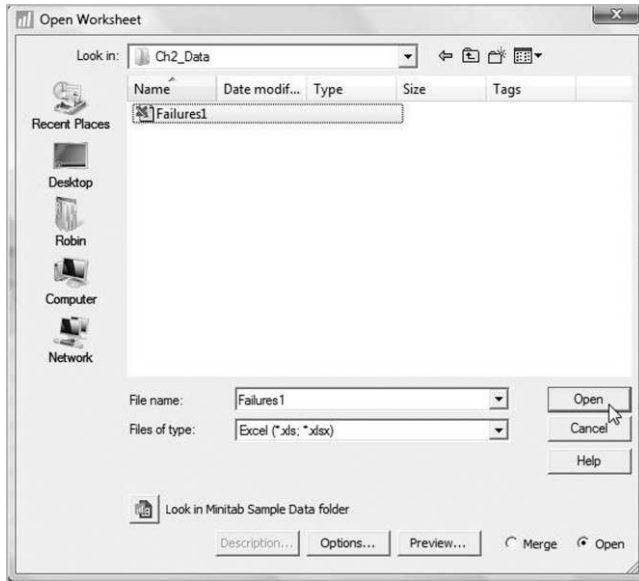


Figure 2.39 Dialog for opening an Excel workbook as a Minitab worksheet.

In order to analyse the data in Minitab the first step is to use **File > Open Worksheet...** As indicated in Figure 2.39, use **Files of type:** to select **Excel(*.xls; *.xlsx)**, and use **File name:** to select the required file. Clicking **Open** enters the data into the Data window. Until **Excel(*.xls; *.xlsx)** has been selected you will not see any Excel workbook files listed. (Note the list of file types catered for. Minitab can directly read and write Minitab portable, Excel, Spreadsheet XML, Quattro Pro, Lotus 1-2-3, and dBase files and text files, with extensions.txt or.csv, or data files with extension.dat.)

2.3.2 Stacking and unstacking of data; changing data type and coding

In order to create a run chart of the number of units per day that fail to pass final inspection it is necessary to stack the blocks of values in columns C2 to C7 of the Minitab worksheet that correspond to columns B to G of the Excel worksheet displayed in Figure 2.38 on top of each other so that the daily numbers of failures appear in time order in a single column. This can be achieved using **Data > Stack > Columns...** Note the descriptive icons positioned beside many of the items on the **Data** menu in Figure 2.40.

The columns to be stacked are selected as indicated in Figure 2.41. Selection may be made by highlighting all six columns simultaneously and clicking the **Select** button. The option to **Store stacked data in:New worksheet** was accepted and **Name:** Daily Failures specified for the new worksheet. In addition, the default to **Use variable names in subscript column** was accepted.

On clicking **OK** the new worksheet is created. Column C1 is named Subscripts by the software and column C2, containing the stacked data, is unnamed. Note in Figure 2.42 how Minitab has automatically created the column of subscripts, which are simply the names of the columns containing the failure counts in the original worksheet. Figure 2.42 shows a portion of the new worksheet with column C1 renamed **Week** and column C2 named

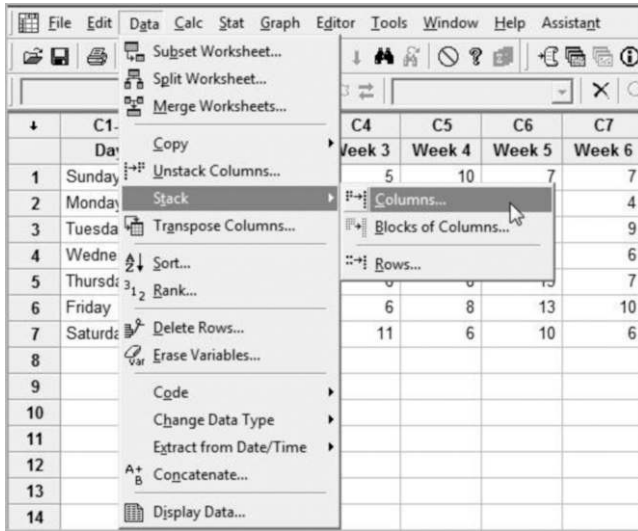


Figure 2.40 Selecting Stack > Columns... from the Data menu.

Failures/Day. (Note that the dialog displayed in Figure 2.41 offers the option of creating the stacked column in the current worksheet, either with or without a column of subscripts.) The data are now arranged in time order and a run chart may be created – this is left as an exercise for the reader.

The Excel workbook Failures2.xls gives the same data in an alternative format as shown in Figure 2.43. Having opened the Excel workbook as a Minitab worksheet, one may create a column with the data in time order in the same worksheet. Use is required of **Data > Stack > Rows...** with dialog as displayed in Figure 2.44. **Store stacked data in:** C12 indicates that the stacked data are to be stored in column C12. The option **Store row subscripts in:** was checked

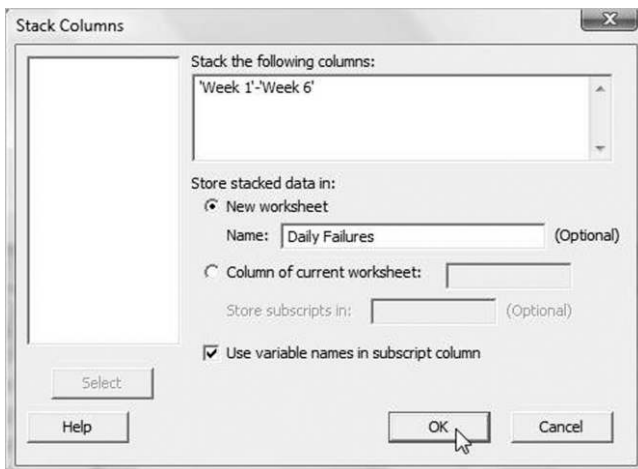


Figure 2.41 Dialog for Stack Columns.

↓	C1-T	C2
	Week	Failures / Day
1	Week 1	4
2	Week 1	9
3	Week 1	7
4	Week 1	10
5	Week 1	7
6	Week 1	13
7	Week 1	11
8	Week 2	3
9	Week 2	7
10	Week 2	12
11	Week 2	5
12	Week 2	7
13	Week 2	11
14	Week 2	17
15	Week 3	5

Figure 2.42 Worksheet containing the stacked data.

	A	B	C	D	E	F	G	H
1	Day	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
2	Week 1	4	9	7	10	7	13	11
3	Week 2	3	7	12	5	7	11	17
4	Week 3	5	6	6	7	6	6	11
5	Week 4	10	9	4	10	8	8	6
6	Week 5	7	7	12	12	13	13	10
7	Week 6	7	4	9	6	7	10	6

Figure 2.43 Alternative layout for failure data.

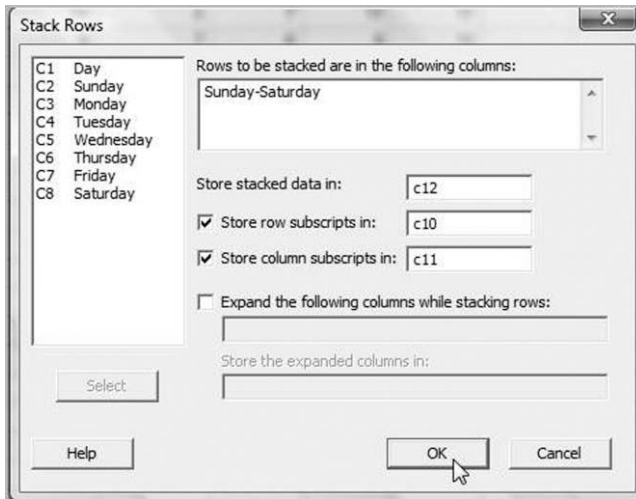


Figure 2.44 Dialog for stacking rows.

and the column C10 specified in the window; the option **Store column subscripts in:** was checked and the column C11 specified in the window.

The new columns require naming. Day is already in use as a column name in the worksheet so the name Day of week was used for the new day column. Column names cannot be duplicated in Minitab. A portion of the stacked data is shown in Figure 2.45. (The allocation of columns for storage of the stacked data and the subscripts in the order C12, C10 and C11 will now appear logical! Column names could have been entered directly during the dialog. Names,

C10	C11-T	C12
Week	Day of week	Failures / day
1	Sunday	4
1	Monday	9
1	Tuesday	7
1	Wednesday	10
1	Thursday	7
1	Friday	13
1	Saturday	11
2	Sunday	3
2	Monday	7
2	Tuesday	12

Figure 2.45 Portion of the stacked data.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Day	Wk1	Wk1	Wk2	Wk2	Wk3	Wk3	Wk4	Wk4	Wk5	Wk5	Wk6	Wk6
2	Line	A	B	A	B	A	B	A	B	A	B	A	B
3	Sunday	2	2	1	2	2	3	4	6	2	5	3	4
4	Monday	3	6	3	4	2	4	3	6	3	4	0	4
5	Tuesday	1	6	4	8	1	5	3	1	5	7	3	6
6	Wednesday	4	6	3	2	3	4	4	6	6	6	2	4
7	Thursday	2	5	2	5	4	2	3	5	5	8	3	4
8	Friday	4	9	4	7	2	4	4	4	4	9	4	6
9	Saturday	3	8	4	13	4	7	2	4	5	5	2	4

Figure 2.46 Failure data stratified by production line.

such as Day of week, that are not simple text strings must be entered enclosed in single quotes. The reader should not be afraid to experiment – if an initial attempt to achieve an objective fails then try again or seek assistance via Help or StatGuide.) Again the data are now arranged in time order and a run chart may be created.

Suppose that production actually involves two lines, A and B, and that the data stratifies by line as shown in the Excel worksheet displayed in Figure 2.46. The data are available in the Excel workbook Failures3.xls.

On opening the workbook as a Minitab worksheet the data appear as shown in Figure 2.47. Note how the software names the two columns labelled Wk1 in the Excel spreadsheet as Wk1 and Wk1_1 in order to have unique Minitab column names. We are faced with the problem of having the first row containing the text values A and B and that therefore Minitab ‘sees’ the columns containing the numerical data we wish to analyse as text columns. This is indicated by C2-T, C3-T, C4-T etc.

The first step in overcoming this is to highlight the entire first row of worksheet entries by clicking on the row number 1 at the left-hand side of the worksheet. On doing this the row will appear as in Figure 2.47. Use of **Edit > Delete Cells** will delete the entire row of unwanted text entries. The next step involves use of the facilities for changing data types available via the **Data** menu. The six types of changes that may be made are indicated in Figure 2.48. Here we require a change of data type from text to numeric. **Data > Change Data Type > Text to Numeric...** gives the dialog box displayed in Figure 2.49. The 12 text columns containing the numerical data are specified in **Change text columns:** and the same column names are specified in **Store numeric columns in:**

Having changed the data type to numeric, C2-T becomes C2 etc. Next we need to use **Data > Stack > Blocks of Columns...** to stack the numeric columns in blocks of two, corresponding to the two production lines, A and B. The completed dialog box is shown in Figure 2.50. Each block of two columns corresponds to one week’s data.

+	C1-T	C2-T	C3-T	C4-T	C5-T	C6-T	C7-T	C8-T	C9-T	C10-T	C11-T	C12-T	C13-T
	Day	Wk1	Wk1_1	Wk2	Wk2_1	Wk3	Wk3_1	Wk4	Wk4_1	Wk5	Wk5_1	Wk6	Wk6_1
1	Line	A	B	A	B	A	B	A	B	A	B	A	B
2	Sunday	2	2	1	2	2	3	4	6	2	5	3	4
3	Monday	3	6	3	4	2	4	3	6	3	4	0	4
4	Tuesday	1	6	4	8	1	5	3	1	5	7	3	6
5	Wednesday	4	6	3	2	3	4	4	6	6	6	2	4
6	Thursday	2	5	2	5	4	2	3	5	5	8	3	4
7	Friday	4	9	4	7	2	4	4	4	4	9	4	6
8	Saturday	3	8	4	13	4	7	2	4	5	5	2	4

Figure 2.47 Stratified data in Minitab.

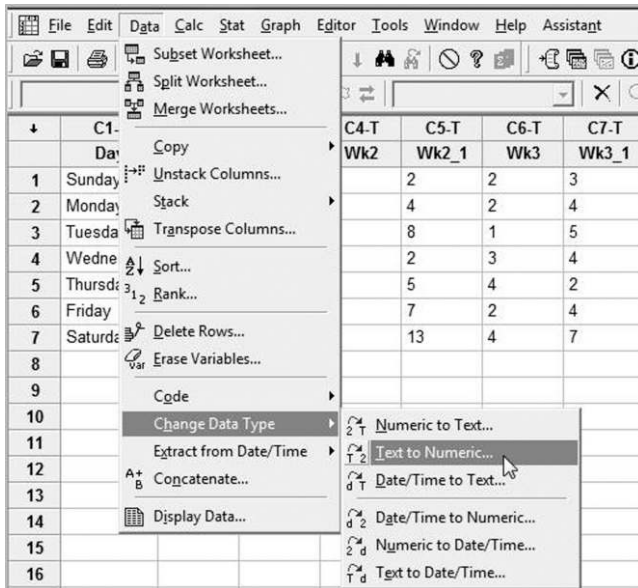


Figure 2.48 Changing data type from text to numeric.

The six blocks of columns to be stacked are specified under **Stack two or more blocks of columns on top of each other:**. In the dialog box shown in Figure 2.50 the default option of storing the stacked data in a new worksheet has been accepted and the name Failures Lines A & B has been specified for it. The default option to **Use variables in subscript column** has also been accepted. This subscript column will appear to the left of a pair of columns, the first of which will contain the sequence of daily failure counts for line A and the second those for line B. These may then be named **Week, Line A** and **Line B** respectively.

It would be informative to see a display of run charts for both lines on the same diagram. One may use **Stat > Time Series > Time Series Plot...** to create such a display. Choose the **Multiple** option, select the two columns containing the data to be plotted in the **Series:**

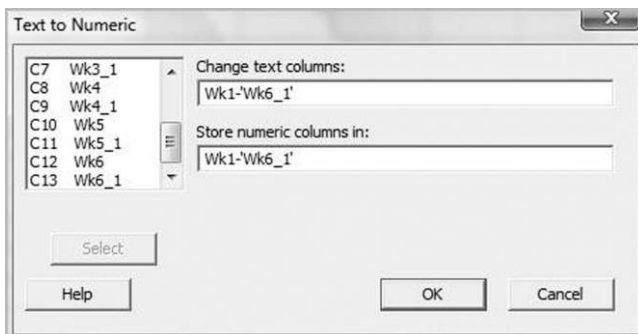


Figure 2.49 Changing text columns to numeric.

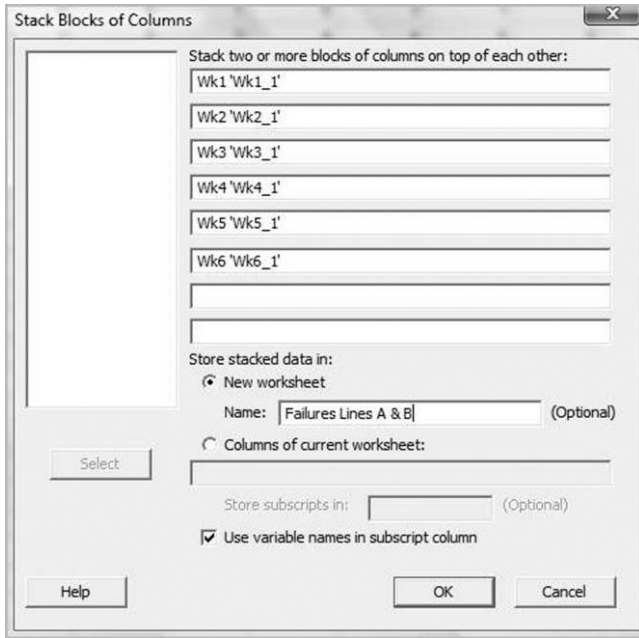


Figure 2.50 Procedure to stack the six blocks of pairs of columns.

window and insert a suitable title in the **Title:** window under **Labels...** Double-clicking on each of the axis labels enables the labels to be edited appropriately. The plot is displayed in Figure 2.51.

Clearly line B has a higher daily rate of failures than line A. You should verify that the medians are 3 and 5 failures per day for lines A and B, respectively. The stratification of the

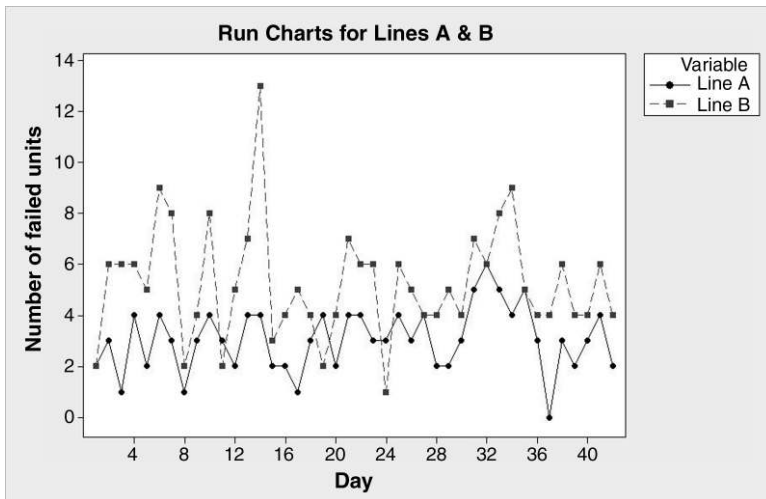


Figure 2.51 Superimposed run charts for lines A & B.

↓	C1	C2	C3	C4	C5	C6	C7	C8
	Pulse1	Pulse2	Ran	Smokes	Sex	Height	Weight	Activity
83	68	68	2	2	2	69.00	150	2
84	72	68	2	2	2	68.00	110	2
85	82	80	2	2	2	63.00	116	1
86	76	76	2	1	2	62.00	108	3
87	87	84	2	2	2	63.00	95	3
88	90	92	2	1	2	64.00	125	1
89	78	80	2	2	2	68.00	133	1
90	68	68	2	2	2	62.00	110	2
91	86	84	2	2	2	67.00	150	3
92	76	76	2	2	2	61.75	108	2
93								

Figure 2.52 Portion of the pulse data set.

daily rate of failures by production line has provided the insight that the performance of line B is worse than that of line A. Given that the lines have identical production capacity, quality improvement could potentially be achieved through investigation of factors contributing to the poorer performance of line B.

A series of data sets that are referred to in examples provided via Help are provided in the Minitab Sample Data folder (typically located in the folder C:\Program Files\Minitab\Minitab16\English). Alternatively, the folder may be accessed by selecting **File > Open Worksheet...** and then clicking on the icon labelled **Look in Minitab Sample Data folder** that appears near the foot of the Open Worksheet dialog box. In order to illustrate further aspects of data manipulation the reader is invited to open the worksheet Pulse.MTW from the Sample Data folder. The worksheet contains data for a group of 92 students. In an introductory statistics class, the group took part in an experiment. Each student recorded their height, weight, gender, smoking habit, usual activity level, and pulse rate at rest. Then they all tossed coins; those whose coins came up heads were asked to run on the spot for a minute. Finally, the entire class recorded their pulse rates for a second time. The data for the final ten students are shown in Figure 2.52.

In this data set codes are used for gender in the column labelled Sex, with 1 representing male and 2 representing female. Before analysing a data set it is always wise to check for any unusual values appearing because of errors in data entry etc. One simple check would be a tally of the values appearing in the Sex column. This can be achieved using **Stat > Tables > Tally Individual Variables...** By default **Counts** are given but the Session window output in Panel 2.4 was achieved by also checking the **Percents** box. Thus there were 92 students in the class, of whom 57 were male, i.e. 62% (to the nearest whole number). Had

Tally for Discrete Variables: Sex		
Sex	Count	Percent
1	57	61.96
2	35	38.04
N=	92	

Panel 2.4 Counts of males and females.

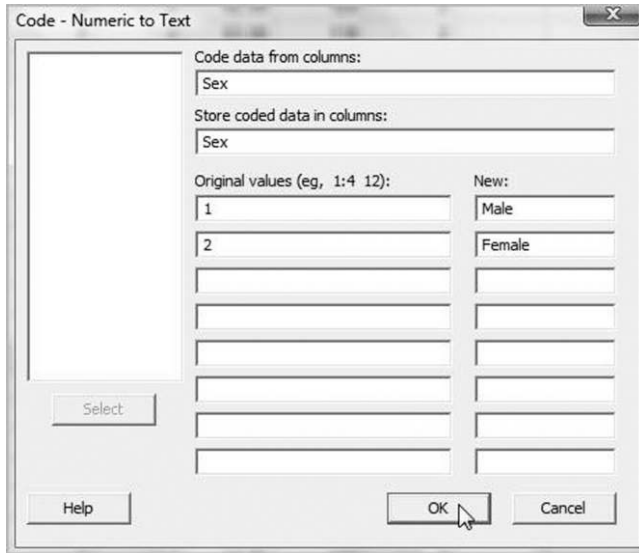


Figure 2.53 Changing numeric codes to text form.

values such as 0 or 3 for Sex appeared then that would have indicated an error in the data or in the data input.

Suppose we wish to replace the numerical codes 1 and 2 with the words Male and Female respectively in the Sex column. This can be achieved using **Data > Code > Numeric to Text...** One reason for using text rather than numerical values, for example, is that displays of the data can be created having more user-friendly labels. The dialog is shown in Figure 2.53. **Code data from columns:** Sex and **Store coded data in columns:** Sex means that the coded text values will be stored in the same column of the worksheet as the original numerical codes. Under **Original values:** note that 1 and 2 have been entered, with the corresponding replacement codes of Male and Female respectively specified under **New:**

If a separate worksheet is required of the data for females then this can be achieved using **Data > Unstack Columns...** All eight columns were selected as the source of the data to be unstacked in **Unstack the data in:** (This may be done by highlighting all eight variables and clicking the **Select** button.) The subscripts to be used for unstacking are in the Sex column and this is indicated via **Using subscripts in:** In the dialog box in Figure 2.54, the default option **Store unstacked data in a new Worksheet** has been accepted with **Name:** Females specified. The default option **Name the columns containing the unstacked data** was accepted. Thus names would automatically be assigned to the columns containing the unstacked data.

The first eight columns of the new worksheet contain the data for the females while the second eight columns contain the data for the males. Note the column headings such as Pulse1_Female in the new worksheet. Minitab has used the subscripts employed for unstacking the original data to extend the original column names appropriately. To delete the data for males from the new worksheet, **Data > Erase Variables...** may be used, with all names ending in _Male being selected. (Alternatively a left click on the cell containing the text C9, keeping the mouse button depressed and scrolling across to the cell containing the text C16

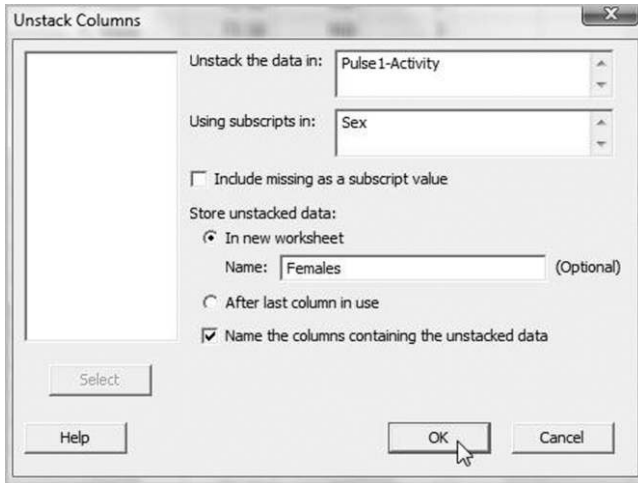


Figure 2.54 Unstacking columns to separate the male and female data.

leads to all the unwanted columns being blacked out. Releasing the mouse button and pressing the delete key completes the operation.) The redundant fifth column may also be deleted by clicking on the cell containing C5-T and using **Edit > Delete Cells**. Finally, double-clicking each of the remaining column names in turn enables them to be edited to their original form as shown in Figure 2.55.

Suppose that it is required to have the columns in the order Height, Weight, Activity, Smokes, Pulse1, Ran and Pulse2. The adjacent Height and Weight columns may be moved first as follows. Click on C5 at the head of the Height column, keep the mouse button depressed and drag across to C6 so that the Height and Weight columns are highlighted as shown in Figure 2.56. Choose **Editor > Move Columns...**, check **Before column C1** and click **OK**.

Further use of the facility for moving columns yields the worksheet in the desired format. The worksheet may then be saved using **File > Save Current Worksheet...** Under **Save as type:** the default is Minitab. If this option were selected then the worksheet could be saved, for example, as **Females.MTW**. (The reader should note the other available file types and observe, when using Windows Explorer, the subtle difference between the icons used for Minitab project and worksheet files.)

In this section a number of methods of data acquisition in Minitab have been considered. There are situations where the capture of data in real time is of interest. Though Minitab was

Females ***							
↓	C1	C2	C3	C4	C5	C6	C7
	Pulse1	Pulse2	Ran	Smokes	Height	Weight	Activity
1	96	140	1	2	61.00	140	2
2	62	100	1	2	66.00	120	2
3	78	104	1	1	68.00	130	2
4	82	100	1	2	68.00	138	2
5	100	115	1	1	63.00	121	2

Figure 2.55 Section of the data for females.

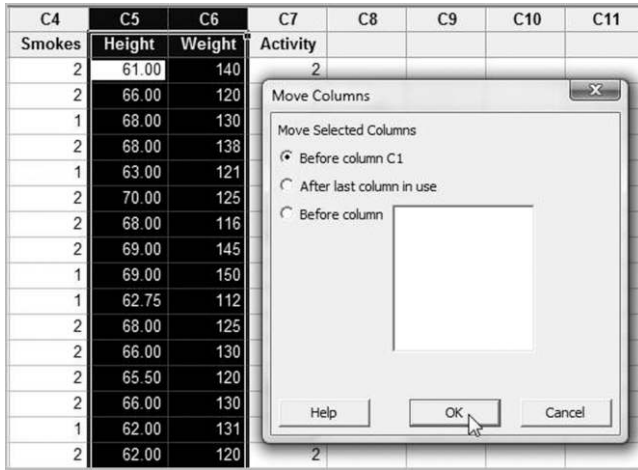


Figure 2.56 Moving columns.

not designed to capture data in real time, software is available to acquire data in real time from devices and transmit it to Minitab. Further information may be obtained from the Minitab website (<http://www.minitab.com/en-GB/support/answers/answer.aspx?id=918>).

2.3.3 Case study demonstrating ranking, sorting and extraction of information from date/time data

As a final example in this chapter we will consider data for patients referred for a colposcopy at a major hospital. The data cover the period from October 2001 to May 2003. A clinical improvement project was commenced in September 2002 in order to improve waiting times for the patients. The data are stored in the supplied Excel spreadsheet *Colposcopy.xls* and are reproduced by permission of NHS Lothian and the Colposcopy Services Clinical Improvement Project Team at the Royal Infirmary of Edinburgh, led by Sister Audrey Burnside. On opening the file as a Minitab worksheet it appears as displayed in Figure 2.57.

The first column is the patient reference number, the second gives the date on which the need for an outpatient appointment for a colposcopy was established and the third gives the date on which the procedure was actually carried out. Note that Minitab has recognized the data in the second and third columns as dates; this is indicated by the column headings C2-D and C3-D. The * symbol in the third column for the seventh patient is the missing value code for numeric data in Minitab – the corresponding cell in the Excel spreadsheet is blank. The aim of the example is to demonstrate how to create a run chart of monthly means of waiting times to indicate process performance before and after process changes. Before creating the run chart, screening of the data for anomalous values will be carried out. This process introduces a number of important facilities in Minitab for the manipulation of data.

2.3.3.1 Step 1. Calculation of waiting times

Calculations may be performed on data in the form of dates. Use **Calc > Calculator...** to calculate the waiting time in days for each patient as indicated in Figure 2.58. **Assign as a**

↓	C1	C2-D	C3-D
	Patient No.	Referral Date	Colposcopy Date
1	1	01/10/2001	26/10/2001
2	2	01/10/2001	03/01/2002
3	3	01/10/2001	08/01/2002
4	4	01/10/2001	20/11/2001
5	5	01/10/2001	20/12/2001
6	6	01/10/2001	29/01/2002
7	7	01/10/2001	*
8	8	01/10/2001	18/02/2002
9	9	01/10/2001	11/12/2001
10	10	01/10/2001	06/12/2001
11	11	01/10/2001	24/01/2002
12	12	01/10/2001	04/02/2002

Figure 2.57 Portion of the colposcopy data.

formula was checked – the reason will be explained below. Note how the waiting time for the first patient was 25 days and that the waiting time for the seventh patient is, of course, a missing value. Note, too, the small green cross at the head of the column of Wait values, indicating that the formula used in the calculation has been assigned to the cells in the column. This means that should any dates in the second or third columns be changed, Wait will be automatically



Figure 2.58 Calculating patient waiting time.

recalculated as in spreadsheet software. In addition, if dates for further patients were to be added then their Wait values would be calculated automatically.

2.3.3.2 Step2. Screening the data for anomalous values

Firstly, use can be made of **Data > Rank...** to rank the values of Wait from lowest to highest. In the dialog box enter **Rank data in:** Wait and **Store ranks in:** Rank. Note that the Wait of 25 days for the first patient is ranked 379. When two or more patients have the same value for Wait then the mean of the corresponding ranks is allocated as Rank for these patients.

Next, use can be made of **Data > Sort...** to sort the waiting times in ascending order and to store the sorted data in a second worksheet. The dialog is shown in Figure 2.59. Note that all five columns should be selected in the **Sort columns(s):** window and that sorting **By column:** Rank is specified. Sorting in ascending order is the default so **Descending** is left unchecked. **Store sorted data in:** Ordered data was used to name the new worksheet. The first few rows of the sorted data are shown in Figure 2.60.

Scrutiny of the ranked data reveals errors. For example the lowest Wait value, with Rank value 1, was -4 . This is impossible, so the data for patient number 1725 require checking. There were 19 patients with Wait values 0, which means that they underwent the procedure on the day that it was deemed necessary – an occurrence likely in situations where clinicians suspected a serious situation for the patient. These 19 patients would be assigned rank values ranging from 2 to 20 inclusive. These values sum to 209, which on division by 19 yields 11. Thus the rank assigned to each of these 19 patients is 11. The 21st and 22nd ordered Wait values were both 1 day, so the corresponding patients are assigned rank 21.5.

The final section of the sorted data is shown in Figure 2.61. Further relevant information emerges. Five patients have no dates for the procedure recorded. The patient with reference number 1964 had a wait of 1122 days, which exceeds the length of the study period. Suppose that discussion with the project leader reveals that the colonoscopy dates for patients 1725 and

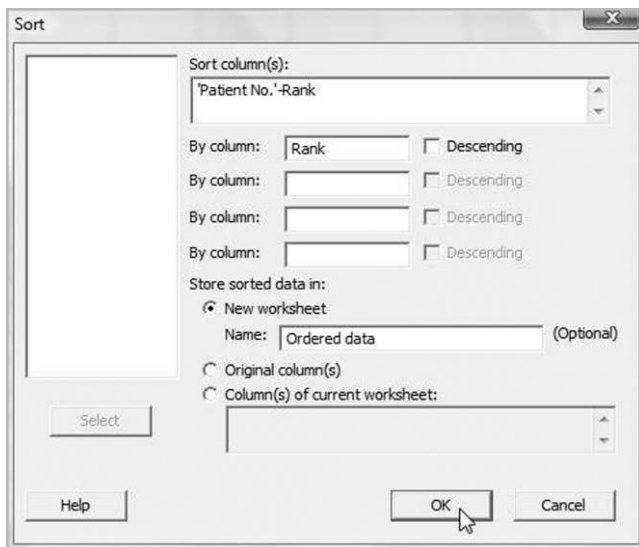


Figure 2.59 Ordering the Wait values.

Patient No.	Referral Date	Colposcopy Date	Wait	Rank
1725	24/01/2003	20/01/2003	-4	1.0
38	04/10/2001	04/10/2001	0	11.0
262	27/11/2001	27/11/2001	0	11.0
434	29/01/2002	29/01/2002	0	11.0
575	11/03/2002	11/03/2002	0	11.0
577	14/03/2002	14/03/2002	0	11.0
789	14/05/2002	14/05/2002	0	11.0
904	06/06/2002	06/06/2002	0	11.0
965	21/06/2002	21/06/2002	0	11.0
1235	23/08/2002	23/08/2002	0	11.0
1348	23/09/2002	23/09/2002	0	11.0
1427	24/10/2002	24/10/2002	0	11.0
1460	04/11/2002	04/11/2002	0	11.0
1530	25/11/2002	25/11/2002	0	11.0
1585	09/12/2002	09/12/2002	0	11.0
1600	10/12/2002	10/12/2002	0	11.0
1708	16/01/2003	16/01/2003	0	11.0
1939	31/03/2003	31/03/2003	0	11.0
2041	28/04/2003	28/04/2003	0	11.0
2042	28/04/2003	28/04/2003	0	11.0
105	17/10/2001	18/10/2001	1	21.5
246	22/11/2001	23/11/2001	1	21.5
153	29/10/2001	01/11/2001	3	25.0

Figure 2.60 Initial section of the sorted data.

1964 were 20/02/2003 and 29/04/2003, respectively. Suppose, too, that she indicates that the patients with reference numbers 7, 1238, 1764 and 1931 should be removed from the data set, as they had moved away from the area served by the hospital. There may still, of course, be further errors. The worksheet of ordered data may be deleted and the appropriate corrections and deletions made in the original worksheet.

Ordered Data ***					
↓	C1	C2-D	C3-D	C4	C5
	Patient No.	Referral Date	Colposcopy Date	Wait	Rank
2126	1015	03/07/2002	28/02/2003	240	2126.0
2127	1770	10/02/2003	17/12/2003	310	2127.0
2128	229	19/11/2001	17/10/2002	332	2128.0
2129	93	15/10/2001	24/09/2002	344	2129.0
2130	480	10/02/2002	11/02/2003	366	2130.0
2131	1430	24/10/2002	04/11/2003	376	2131.0
2132	1168	08/08/2002	11/11/2003	460	2132.0
2133	1964	03/04/2003	29/04/2006	1122	2133.0
2134	7	01/10/2001	*	*	*
2135	1238	23/08/2002	*	*	*
2136	1764	07/02/2003	*	*	*
2137	1931	21/03/2003	*	*	*

Figure 2.61 Final section of sorted data.

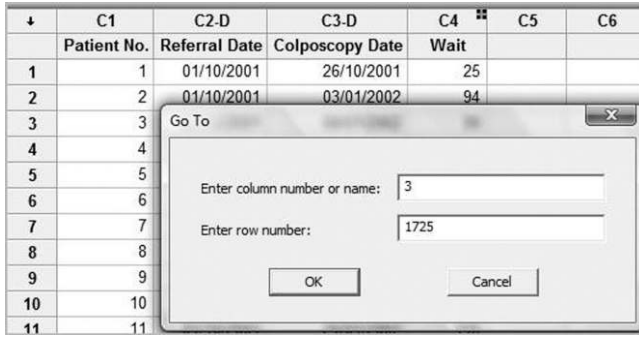


Figure 2.62 Locating a cell for correction.

2.3.3.3 Step 3. Correcting the data

The date corrections can be made first. One can scroll around the worksheet in order to locate the cells requiring to be changed or alternatively use **Editor > Go To...** (the command with an icon consisting of two footprints) when the worksheet is active. An example of the dialog involved in locating a cell for correction is shown in Figure 2.62 in the case of the patient with reference 1725 whose referral date should be 20/02/2003. Double-clicking the relevant cells enables the edits to be made. Observe how the Wait value for this patient changes automatically from -4 to 27 on making the correction. The change for the patient with number 1964 may be made similarly.

In order to make the deletions use may be made of **Data > Delete Rows...** to specify the rows to be deleted, as shown in Figure 2.63. Note that in this case the patient number matches the row number and that the rows have to be deleted from all four of the columns available for selection. The Wait column is not available for selection as it was assigned a formula. The now redundant column Rank may also be deleted using **Data > Erase Variables...**

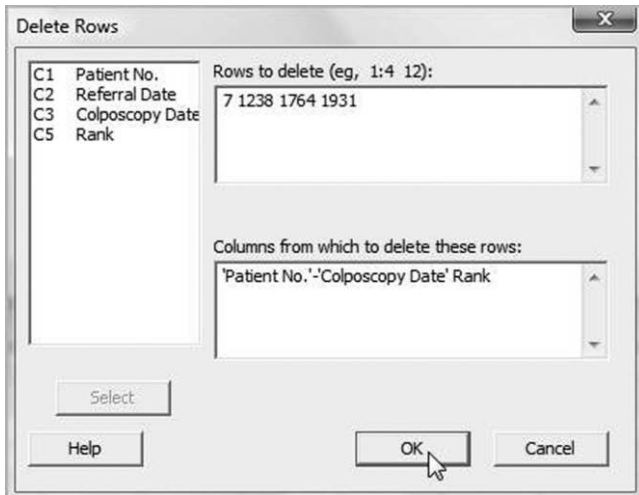


Figure 2.63 Deleting rows.

Descriptive Statistics: Wait									
Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Wait	2133	0	65.272	0.951	43.935	0.000	31.000	56.000	96.000
Variable	Maximum								
Wait	460.000								

Panel 2.5 Descriptive statistics for Wait.

For readers who are following the steps ‘live’, a crosscheck at this point can be made by obtaining descriptive statistics for the revised column of waiting times as shown in Panel 2.5. Note that there are 2133 values of Wait, with none missing, and that the mean was 65.272 days.

2.3.3.4 Step 4. Grouping the patients by month appointment was made

Use can be made of **Data > Extract from Date/Time > To Numeric** to convert the full date a patient’s referral was made to a code for the month during which the referral was made. The dialog is displayed in Figure 2.64.

Under **Specify at least one component to extract from date/time** the **Year** was checked, with the **Four Digit** option selected, and **Month** was also checked. With the selection of the four-digit Year component and the Month component in the Minitab dialog box, any referral date in October 2001 will be coded as 200110, any referral date in November 2001 as 200111 etc. Use of **Stat > Basic Statistics > Descriptive Statistics...** for Wait, with **By variables: Month** and with **Mean** checked under **Statistics...**, as the only statistic required, yields the monthly means in the Session window as in Panel 2.6.

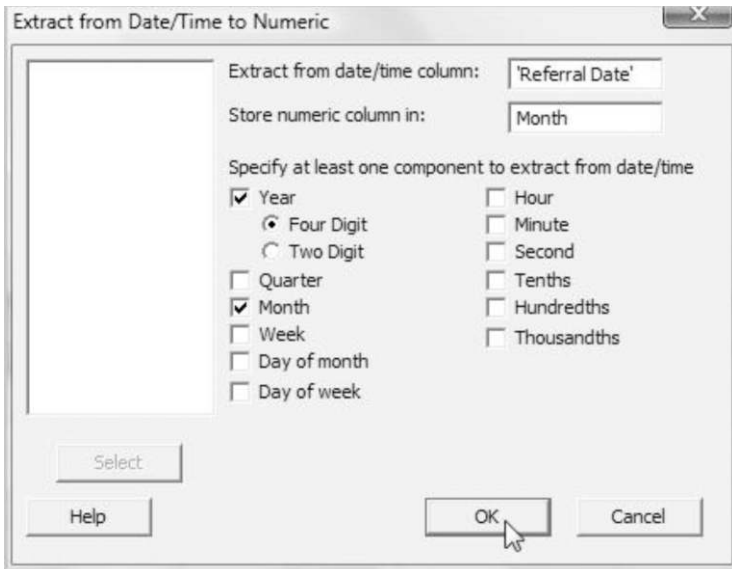


Figure 2.64 Coding dates by month.

Descriptive Statistics: Wait		
Variable	Month	Mean
Wait	200110	80.85
	200111	79.68
	200112	77.69
	200201	87.22
	200202	77.98
	200203	64.32
	200204	72.84
	200205	75.31
	200206	75.38
	200207	78.91
	200208	87.16
	200209	75.61
	200210	58.42
	200211	47.48
	200212	53.74
	200301	38.17
	200302	37.07
	200303	34.21
	200304	36.60
	200305	31.80

Panel 2.6 Mean wait by month of referral.

The means can then be copied from the Session window, along with the code for the months, and pasted into a new worksheet as follows. Having highlighted and copied the two columns of 20 numbers from the Session window, click **File > New...**, and with **Minitab Worksheet** highlighted, click **OK**. With the mouse pointer, click on the first cell in column C1 of the new worksheet before pasting, accepting the default setting to **Use spaces as delimiters**. The run chart displayed in Figure 2.65 may then be created. It would appear that there has been

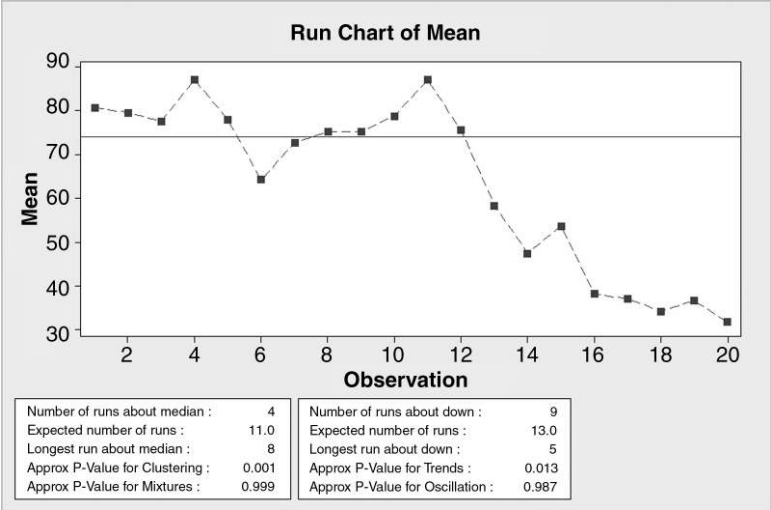


Figure 2.65 Run chart of mean wait by month.

a dramatic reduction in waiting times for patients. Note that the P -values provide evidence of the presence of special cause variation.

The author has heard a run chart described as a naked control chart. In Chapter 5 the construction and use of control charts will be introduced. Chapter 4 will be devoted to the introduction of the basic concepts of probability and of statistical models that provides essential underpinning for control charts.

2.4 Exercises and follow-up activities

1. For a familiar process obtain a sequence of measurements and display them in the form of a run chart. If you do not have access to data from a workplace situation then journey durations, your golf scores etc. could be used. Do you consider the process to be behaving in a stable predictable manner? Use ReportPad to note any conclusions you make regarding process performance. Save your work as a Minitab project to facilitate updating the data and to provide a control charting exercise at a later date. You might wish to name the project file Ch2Ex1.MPJ.
2. The Minitab worksheet `Scottish_Mean_Annual_Temperatures.MTW` gives mean annual Scottish temperatures (degrees Celsius) for the years 1910–2010 (<http://www.metoffice.gov.uk/climate/uk/datasets/Tmean/date/Scotland.txt>, accessed 30 January 2011.) Display the data in the form of a run chart and comment.
3. In the bottle weight example used in this chapter the sample (subgroup) size was 1. The worksheet `Bottles.MTW` contains data giving the weights of samples of four bottles taken every 15 minutes from a bottle-forming process. Open the worksheet and create a run chart. Here the data are arranged as subgroups across rows of columns C2, C3, C4 and C5, so this option has to be selected in the Run Chart dialog.

The default is to plot the means of the subgroups of four weights with reference line placed at the median of the 25 sample means, i.e. 489.638. The means are plotted as red squares and connected by line segments, and in addition the individual weights are plotted as black dots. The alternative option is to plot the subgroup medians, in which case the reference line is placed at the median of the 25 medians, i.e. 489.595.

Note from the run chart legends that there is no evidence of any special cause variation. Stack the weights into a single column named `weight` and verify that the mean and standard deviation of the total sample of 100 bottles are 489.75 and 2.09, respectively. Create a histogram of the data with reference lines placed at the specification limits of 485.0 and 495.0 g for bottle weight. Comment on the shape of the distribution and on process performance in relation to the specifications. Save your work as a Minitab project.

4. During the measure phase of a Six Sigma project a building society collected data on the time taken (measured as working days rounded to the nearest day) to process mortgage loan applications. The data are stored in the supplied worksheet `Loans1.MTW`. Display and summarize the data and comment on the shape of the distribution. Use a stem-and-leaf display to determine the number of times which exceeded the industry standard of 14 working days.

Following completion of the improvement project a further sample of processing times was collected. The data are stored in Loans2.MTW. Display and summarize the data in order to assess the effectiveness of the project. Save your work as a Minitab project.

5. The file Statin.xls contains monthly numbers of stroke patients admitted to a major hospital for the years 2001–2003 together with the numbers whose medication on admission included a statin. Open the file in Minitab and create a column giving the monthly proportions of stroke patients on a statin at time of admission. You will find that **Data > Transpose Columns...** is useful here. Create a run chart of the proportions and comment. Save your work as a Minitab project.
6. The Minitab worksheet Shareprice.MTW supplied with the software in the Data folder contains monthly share prices for two companies ABC and XYZ. Use **Graph > Time Series Plot...** to create a multiple run chart of these data. Select the option **Multiple**, and enter ABC and XYZ in the **Series:** window. Under **Time/Scale...** select **Calendar** and then choose **Month Year** from the associated menu. For the **Start Values**, accept the default **One set for all variables** and enter 1 for Month and 2009 for Year. Observe how moving the mouse pointer to a point on a plot leads to display of the variable, its value and the corresponding month and year.
7. The histogram was introduced as a type of bar chart in which there are no gaps between the bars, indicating that the variable being displayed is continuous. One can use the histogram facility in Minitab to display discrete random variables and to emphasize the discrete nature of the data by having gaps between the bars. The supplied worksheet AcuteMI.MTW contains daily counts of the number of patients admitted to the accident and emergency department of a major city hospital with a diagnosis of acute myocardial infarction. Create a histogram of the data, making use of **Data View...** to uncheck **Bars** and check **Project lines**. Double-click on one of the project lines, select **Custom** and increase **Size** to, say, 5.