



M248
Analysing data
Computer Book A

Introduction

This part of the book covers all the computer work associated with the units in Book A of M248 *Analysing data*. The computer work has two components: Minitab and animations. Minitab is a data-analysis package; it is introduced in Chapter 1 of this book and is used throughout M248. The animations are designed specifically to help develop your understanding and appreciation of particular statistical concepts. The first animation is used in Chapter 5 and is associated with ideas introduced in Unit 2.

See the M248 website for information on the software supplied, including installation instructions.

1 Introducing Minitab

This chapter is associated with Subsection 3.4 of Unit 1.

In this chapter, the data-analysis software package Minitab is introduced. If you have not yet installed the M248 software on your computer, then do so now. Instructions are given on the M248 website.

The Minitab environment is discussed briefly in Subsection 1.1. Subsection 1.2 will give you your first taste of using Minitab to do statistical analysis in this module, by going through the process of producing a bar chart. You will see how to save your work in Subsection 1.3. The remaining subsections, Subsections 1.4 and 1.5, describe how to print output and how to paste output into a word-processor document. Note that Subsections 1.1–1.3 are best studied in the same session.

1.1 Getting started

A first introduction to Minitab is given in Screencast C1. You might prefer to watch the screencast rather than working through Computer activity 1 which follows it.

Screencast C1 Getting started with Minitab



Computer activity 1 Running Minitab

Run Minitab now: double-click on the **Minitab Statistical Software** icon on your desktop (or select ‘Minitab Statistical Software’ from your list of programs). You will see briefly a panel that shows a Minitab logo, in older versions of Minitab, the panel shows the version being used. Following this, you should see the opening screen, as shown in Figure 1.



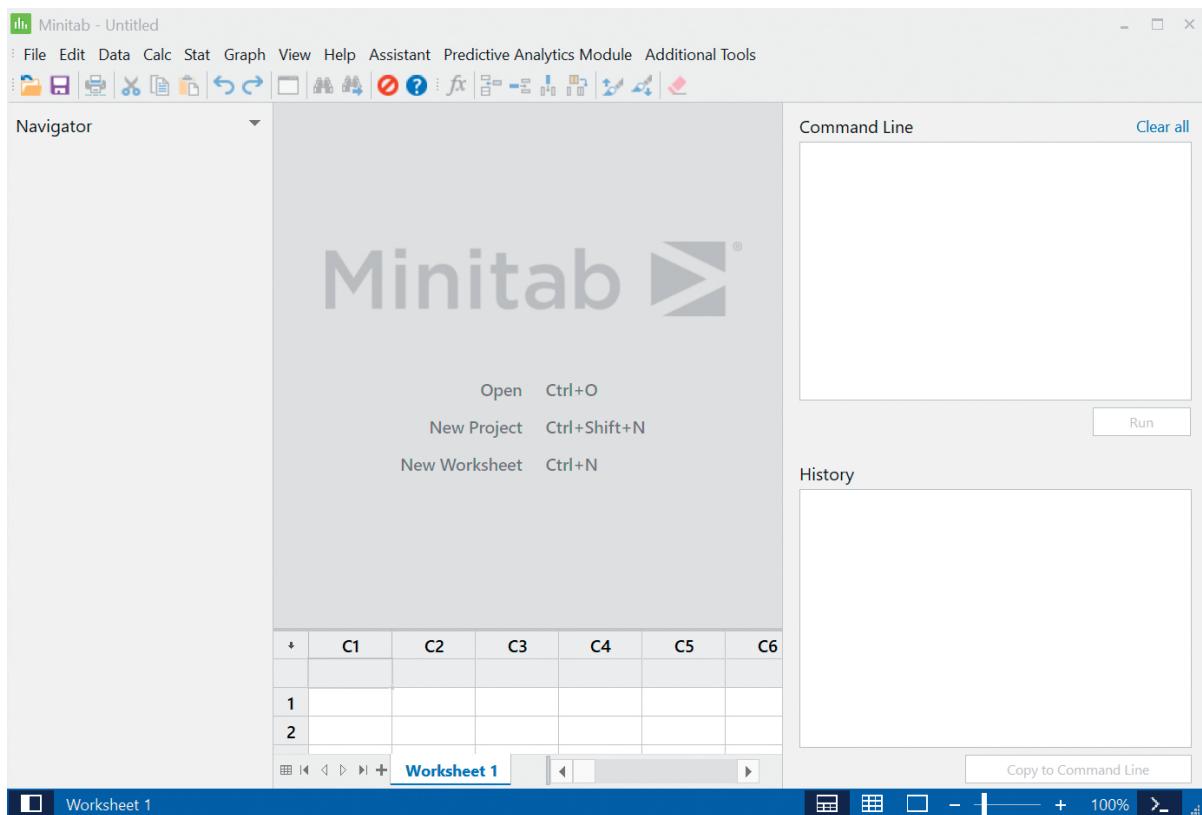


Figure 1 The opening screen in Minitab

The opening screen is similar to that of many Windows-based software packages: there is a menu bar at the top of the screen. Minitab opens with four windows: the top window is called the *Session* window, the bottom one is called the *Data* window, the left-hand one is called the *Navigator* window and the right-hand window (a twin pair of windows) is called *Command Line/History* window. The Session window is where the results are displayed, the Data window is where your data are displayed, the Navigator window is a filing system that handles the results you have produced, and the Command Line/History window is where you can write and run commands and view history of commands you have already run.

Click on **File** in the menu bar at the top to view the contents of the **File** menu. Notice that as you hover over a menu item, a popup box provides information about that item. An arrowhead pointing to the right on a menu item indicates the existence of a submenu. If you click on **File** again, the menu will close.

Before moving on to the next activity, spend a few minutes exploring the menus and their submenus. Note the different types of facilities available in the different menus.

You can exit from Minitab at any time by clicking on the cross at the top right-hand corner of the Minitab window, or by choosing **File > Exit**, which means ‘click on **File** and choose **Exit** (by clicking on it)’. Similar notation will be used throughout the computer book.

The roles of the menus may be broadly summarised as follows.

- The **File** and **Edit** menus contain commands for handling and editing files.
- The **Data** menu allows you to manipulate data in the Data window.
- The **Calc** menu allows you to carry out calculations.
- Statistical techniques are available using the **Stat** menu.
- The **Graph** menu is used to create graphs and diagrams.
- As for other Windows-based software packages, the **View** menu is for rearranging windows or activating a specified window.
- The **Help** menu provides access to online help, and the **Assistant** menu includes help with choosing appropriate techniques for analysing and displaying data. Some of the information is about using Minitab and some about statistics in general – however, much of this is beyond the scope of this module.
- The **Predictive Analytics Module** menu contains additional statistical techniques which are beyond the scope of this module.
- The **Additional Tools** menu provides online access to additional statistical tools and e-learning of statistics. You will not need to use this menu in studying this module either.

In Minitab, data are stored in *worksheets*. When a worksheet is opened, the data it contains are displayed in a Data window. All the worksheets for M248 are located in the **M248 Data Files** folder within **Documents** (or **My Documents**). The filename extension of a Minitab 21 worksheet is **mwx** (**mtw** in versions earlier than Minitab 19). In the next activity, you will explore a Minitab worksheet.

Computer activity 2 Exploring a Minitab worksheet

The data described in Example 6 of Unit 1 on the surgical removal of tattoos are contained in a Minitab worksheet named **tattoos.mwx**. Open this worksheet, as follows.

- Select **File > Open...**. The **Open** dialogue box will open.
- The list of folders in **Documents** will be displayed in the main panel of the **Open** dialogue box. Navigate to the folder **M248 Data Files** where the M248 Minitab worksheets are stored, and then double-click on the folder name to open it.
- Scroll through the list of filenames in the main panel until you find **tattoos.mwx**, then double-click on it. Alternatively, you can open a worksheet by clicking on its name to select it, then on **Open**; or you can type its name in the **File name** field, then click on **Open**.
- The data should now be displayed in the Data window.

A screenshot of the Minitab worksheet **tattoos.mwx** is shown in Figure 2.

	C1-T	C2-T	C3-T	C4-T	C5	C6
	Method	Gender	Size	Depth	Score	
1	A	m	large	deep	1	
2	A	m	large	mod	1	
3	B	f	small	deep	1	
4	B	m	small	mod	4	
5	B	f	large	deep	3	
6	B	m	med	mod	4	
7	B	m	med	deep	4	
8	A	m	large	deep	1	
9	A	m	large	mod	4	
10			small	mod	4	

Figure 2 Screenshot of Minitab worksheet **tattoos.mwx**

This worksheet contains the data for 55 patients: each row of the worksheet contains the data for a single patient. There are five variables: **Method**, **Gender**, **Size**, **Depth** and **Score**. The data for each of these variables is given in the column underneath the variable name. Each of the columns in Minitab is also labelled (by Minitab) above the variable name. Notice that variables **Method**, **Gender**, **Size** and **Depth**, in the first four columns of **tattoos.mwx**, are categorical variables whose possible values are not numerical. Minitab distinguishes the columns corresponding to non-numerical variables from columns corresponding to variables taking numerical values by adding ‘-T’ to the column label. Thus the first four columns are labelled C1-T, ..., C4-T, while the fifth column, which contains numerical data values, is labelled C5.

Minitab contains a description of the data in each of the M248 data files. View a description of the data in the worksheet **tattoos.mwx**, as follows.

- Select **File > Description > Worksheet...**. The **Worksheet Description** dialogue box will open.

A description of the data is contained in the **Comments** field of the **Worksheet Description** dialogue box.

- When you have read the description, click on **OK** (or **Cancel**) to close the **Worksheet Description** dialogue box.

If possible, keep this worksheet open in Minitab for now. You will need it open for Computer activity 3.

The cells in a Data window contain values that you have retrieved by opening a worksheet or that you have typed in directly. The Data window is not a spreadsheet, even though it looks very like one! Cells do not contain formulas: although you can create new columns of values using

values in existing columns, the values in the new columns do not update automatically when values in the existing columns are changed.

Computer activity 3 More windows

You can have several Minitab worksheets open at the same time. In this activity you are going to have two worksheets open at the same time: **tattoos.mwx** and **workforce.mwx**, which contains data on the number of people employed in different occupation types in the UK in 2015.

- Make sure that **tattoos.mwx** is open in Minitab. (This should be the case if you are continuing straight on from Computer activity 2.)
- Using the same procedure as you used in Computer activity 2, open **workforce.mwx**. A reminder of how to do this is given in the margin. (You will find similar reminders given in the margin throughout the computer book.)

A second Data window opens; it is called **workforce.mwx**. This window becomes the active window when the worksheet is opened.

The lowest line of the Data window displays the names of the worksheets you have opened, **tattoos.mwx** and **workforce.mwx**. One of the names is displayed in blue bold-face type. This is the active worksheet, and any commands you give will be applied to it.

- As you have just opened **workforce.mwx**, that is the active worksheet. To make **tattoos.mwx** the active worksheet, click on its filename.

Spend a few moments navigating between Data windows, checking that you can make either worksheet the active one.

The bottom (blue) bar of the Minitab screen contains icons that let you choose which windows are displayed. The icon at the left-hand end of the bar is a toggle-switch for the Navigator window. Click on the icon and the Navigator window disappears; click on it again and the Navigator window reappears. The three icons towards the right-hand end enable you to choose whether to display the Session window and/or the Data window. Click on these icons to see how they work.

The slider at the right of the bar enables you to magnify/reduce the size of the contents of the Session window and Data window. The arrow at the extreme right is a toggle switch for the Command Line/History window. You will not need to use the Command Line/History window in studying this module.

In Minitab, the commands you can access on the **File** menu depend on which window is the active window. For example, when the active window is a Data window, a description of that worksheet can be found via **File > Description > Worksheet...**

These data are described in Example 2 of Unit 1.

File > Description > Worksheet...



Plenty of open windows here ...

If you intend to proceed directly to the next couple of subsections, then do not close the worksheets **tattoos.mwx** and **workforce.mwx**. You will need these worksheets open for the activities in Subsection 1.2.

1.2 Bar charts

In order to explore some of Minitab's features further, it is necessary for you to produce some Minitab output. To this end, in this subsection you will learn how to use Minitab to obtain bar charts similar to some of those described in Subsection 3.1 of Unit 1. You will learn how to save the work done in this subsection in Subsection 1.3, so it would be very helpful to be able to work on both this subsection and the next in a single session.

Bar charts are produced via **Graph > Bar Chart....**. Some of the many options available with **Bar Chart...** will be discussed briefly.

Computer activity 4 Quality of tattoo removal

In this activity you will obtain a bar chart similar to the one in Figure 1 of Unit 1 for the data on the quality of tattoo removal. The data are in the Minitab worksheet **tattoos.mwx**.

File > Open Worksheet...

- Open this worksheet now, if it is not already open, or make sure that **tattoos.mwx** is the current worksheet.

Note that the data are in *raw* form, which means that there is one row of the worksheet for each observation.

- Select **Graph > Bar Chart....** The **Bar Charts** dialogue box will open.

At the top of the dialogue box, there is a field labelled **Bars represent**.

- Click on the arrow to the right of the field to view the **Bars represent** drop-down list.

This list contains three options. **Counts of unique values** (the default) is used when the data are stored in raw form (as is the case here), and **Values from a table** is used when the data are in summary form (as they will be in a later activity). The other option – **A function of a variable** – will not be used in M248.

- The data are in raw form, so select **Counts of unique values**.
- Select a basic bar chart by clicking on the **Simple** diagram.
- Click on **OK**, and the **Bar Chart: Counts of unique values, Simple** dialogue box will open.

Simple is the default option.

To produce a bar chart for the quality of tattoo removal, you will need to enter the name of the column containing the data in the **Categorical variables** field. Any columns of the worksheet that contain data that could be used for a bar chart are listed in the area to the left of the dialogue box – all the columns are listed in this case. Notice that Minitab puts the column labels (without the ‘-T’) next to the variable name; so, for example, the first categorical variable is listed as C1 Method. The categories of interest at the moment correspond to the quality of tattoo removal; this information is given in column C5 Score. (These categories are labelled by the numbers ‘1’ to ‘4’.)

- Enter C5 or Score in the **Categorical variables** field. (The latter is most easily done by double-clicking on C5 Score.)
- Click on **OK**, and the bar chart in Figure 3 will be produced.

We will use column names, like **Score**, rather than column numbers, like C5, from here on.

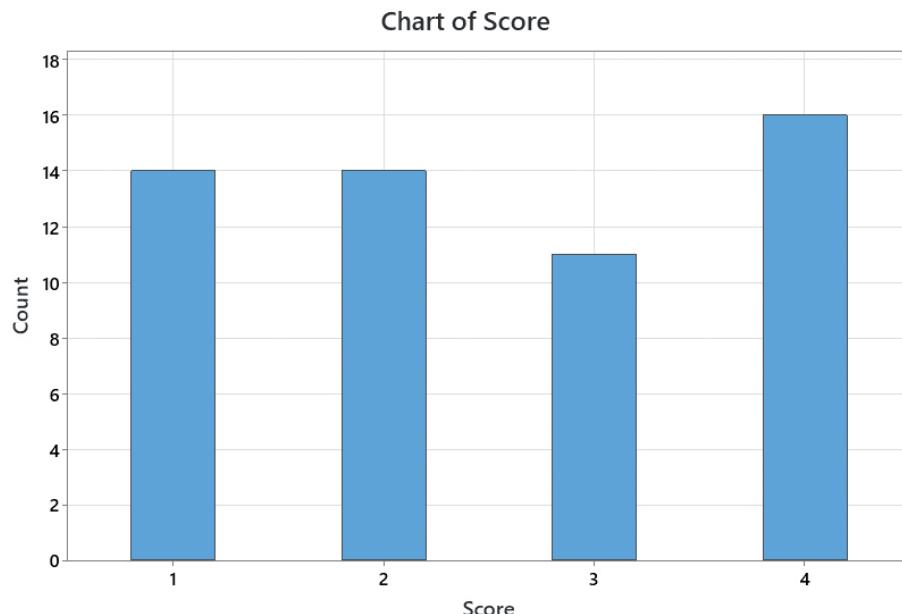


Figure 3 A bar chart of quality of tattoo removal

In the next activity, you will be editing this bar chart. So, if possible, keep this bar chart open in Minitab and move straight on to Computer activity 5.

Computer activity 5 Editing graphs

Figure 4 is simply Figure 3 with the title removed and the label on the vertical axis changed.

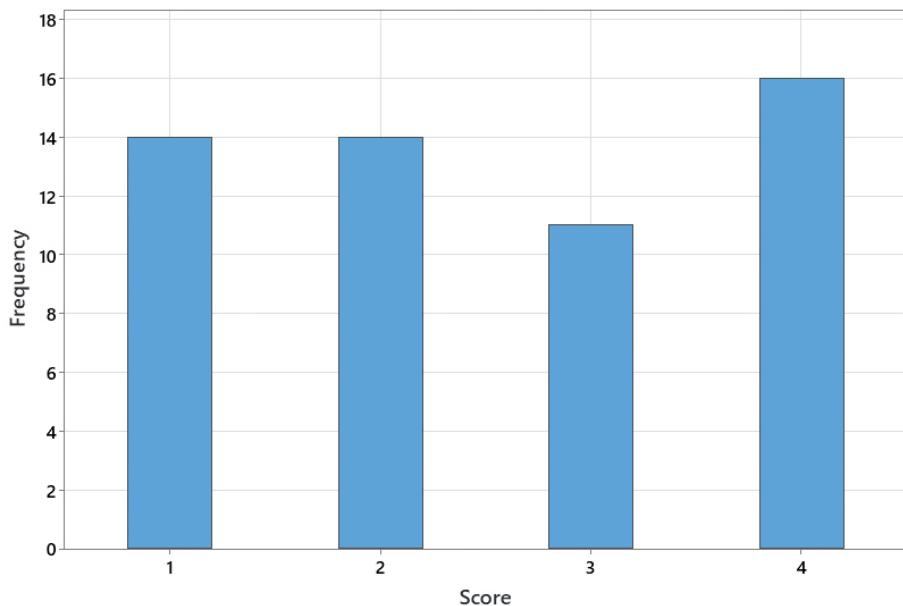


Figure 4 An edited bar chart of quality of tattoo removal

To produce Figure 4, you first need to have Figure 3 as produced in Computer activity 4, then you need to edit the default vertical axis label, as follows.

- Double-click on the bar chart produced in Computer activity 4 to bring up the **Edit Graph** panel.
- Select the label on the vertical axis (by clicking on it), then double-click on it (or press **Ctrl+T**) to open the **Edit Axis Label** dialogue box.
- Replace the default label by typing **Frequency** in the **Text** field at the bottom of the **Edit Axis Label** dialogue box.
- Click on **OK**, and the title of the vertical axis will be replaced.



Labelled axes

The title for a plot can be edited in a similar way (by selecting the title and double-clicking to open the **Edit Title** dialogue box). However, often a title for a plot is not required at all, and indeed many of the default titles for Minitab plots aren't very informative. Titles can be deleted by doing the following.

- Select the title (by clicking on it in the **Edit Graph** panel).
- Press the **Delete** key.

You should now have the bar chart of Figure 4. The Minitab plots shown in the M248 computer book will usually have been edited to delete titles without comment, but it is fine for you to either leave the titles there or delete them as you prefer.

Computer activity 6 Total UK workforce

In this activity, you will obtain a bar chart for the data on the total UK workforce similar to the one in Figure 3 of Unit 1. The data are in the Minitab worksheet **workforce.mwx**. Open this worksheet now if it is not already open, or make sure that **workforce.mwx** is the current worksheet.

There are several points to note here. First, the data are stored in *summary* form in the worksheet. That is, the rows of the worksheet correspond not to individual observations, as they do for raw data, but to groups: in this case, each row of the worksheet corresponds to an occupation type, and the numbers in that row correspond to total numbers of observations falling within the group. Second, the bars in Figure 3 of Unit 1 are arranged in order of length so that the longest bar is first (to the left) and the shortest is last (to the right). Third, the labelling on the axes needs to be changed from the default to reproduce Figure 3 of Unit 1.

- Obtain the **Bar Charts** dialogue box.
- The data are in summary form, so choose **Values from a table** from the **Bars represent** drop-down list.
- Select **Simple** (the default), and click on **OK**. The **Bar Chart: Values from a table, One column of values, Simple** dialogue box will open.

To display data that are in summary form, you must specify the column that contains the frequencies in the **Graph variables** field and the column containing the categories in the **Categorical variable** field.

- Enter **Total** in the **Graph variables** field and **Occupation type** in the **Categorical variable** field. Note that because **Occupation type** is two words, if typing the variable name into the **Categorical variable** field, you will need to type '**Occupation type**' (including the quotation marks) so that Minitab knows that this is the name of a single variable (otherwise, Minitab will look for a variable **Occupation** and will give an error message).
- To specify the order of the bars, click on the **Chart Options...** button to open the **Bar Chart: Options** dialogue box.

It's just as well that the rows don't correspond to individual observations in this case: there would need to be well over 31 million rows if these data were in raw form!

Graph > Bar Chart...

The default ordering displays the categories in numerical order when the categories are given in numerical form or, for text columns, in the order that they occur in the worksheet.

- Select **Decreasing Y**. With this option selected, the longest bar is drawn first and the shortest last.
- Click on **OK** to close the dialogue box and on **OK** again to produce the bar chart.

Finally, delete the default title and change the label on the vertical axis to **Frequency (millions)**. The bar chart will then be as shown in Figure 5.

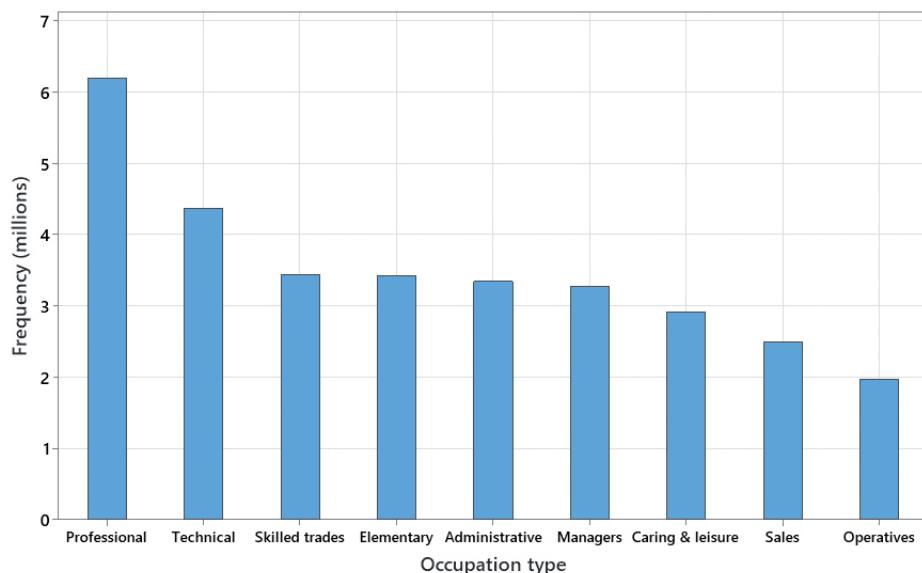


Figure 5 A bar chart of total UK workforce

In the next (short) subsection you will find out how to save the Minitab work that you have done in this subsection, so, if possible, don't exit Minitab yet!

1.3 Saving your work

One of the most frustrating things about using software packages is that if you have to quit a session before you have finished, many of them require you to start afresh the next time you run the package. This is not the case with Minitab. You can save your session in a *project file* and then pick up from where you left off whenever you choose by opening the file.

When you begin a Minitab session, a new project file is opened automatically, so you do not have to remember to open a project file at the beginning of a session.

Computer activity 7 Saving your session in a project file

Use the following instructions to save your current session in a project file named **project1.mpx**.

- If you are following straight on from Subsection 1.2, then you should have two worksheets open, **tattoos.mwx** and **workforce.mwx**, together with two bar charts. If not, then open the two worksheets now. It would also be helpful to have the bar charts from Computer activities 5 and 6 displayed.
- Choose **File > Save Project**. The **Save Project As** dialogue box will open.
- If necessary, navigate to the **M248 Data Files** folder or another folder where you want to save your M248 computer work.
- Type **project1** in the **File name** field.
- Click on **Save** or press **Enter**.

Minitab 19 project files have the filename extension **mpx** (**mpj** in earlier Minitab versions).

Minitab adds **.mpx** (or **.MPX**, depending on your system) automatically.

At any time, you can view a session that you have saved as a project by opening the project file. Choose **File > Open...** to obtain the **Open** dialogue box, navigate to the folder where the project file is located, and double-click on the filename. You will find that your project is restored exactly as you left it, complete with worksheets, graphs, etc., and you can carry on from where you left off. If, after further work, you wish to save your session under a different filename, choose **File > Save Project As...**. You will be offered the chance to choose another filename. Choosing **Save Project** instead will, of course, overwrite the old file with the modified one.

Note that when you choose **File > Exit**, you will be asked whether or not you wish to save changes to your project before closing Minitab.

The contents of an individual output window may be saved by making the window active and clicking on **File**, and then choosing **Save Current Output As...**. The window contents will be saved as an HTML file, with extension **htm** or **html**.

The worksheets can be saved in a similar way. The only difference is that the dialogue boxes are obtained by clicking on **Save Worksheet As...** and the worksheet must be the active window.

A graph can also be saved with the extension **png**, **jpg**, **tif**, **bmp**, **gif** or **emf**. To do this, make the graph active, click on the small box with an arrow, and select **Save Graph As...** from the drop-down menu.

If you wish to save a project or graph (or whatever) in a different folder from that containing the M248 Minitab data files, then navigate to the directory you want to use, and click on **Save**.

By the way, you can edit the contents of the Session window. This means that you can remove unwanted text and annotate your output, if you wish, before saving it.

Having saved your session, now is the time to take a break if you need to. Although we will briefly return to the worksheets that you used and the bar charts that you created in Subsection 1.2, this can be done using the project file that you have just saved in Computer activity 7.

1.4 Printing output

The Navigator window lists the pieces of output that you have. You can print one of them by doing the following.

- In the Navigator window, click on the output you want to print.
- In the Session window, click once on the output. A gray border will appear around it, showing what will be printed.
- Also, a small box containing a downward-pointing arrow will appear just outside the top-right corner of the gray border. Click on this and a list of options will appear. Click on **Print**....
- The **Print** dialogue box then opens, allowing you to choose your printer options. Make sure that **Print range** (or **Page range**, depending on what you are printing) is set to **Selection**, when this option is available. Check that the other options are correct before clicking on **OK**.
- Alternatively, to print the entire contents of the Session window, make sure that **Print range** is set to **All**.
- You can also bring up the **Print** dialogue box by choosing **File > Print**.

Although printing output in this way is simple, it has the disadvantage that it can use a whole sheet of paper for one comparatively small picture or a small amount of text. If you are accustomed to word processing, you may want to paste your output into a word-processor document and then print the document. Pasting output into a word-processor document is described in Subsection 1.5.

1.5 Pasting output into a word-processor document

Incorporating output from Minitab into a word-processor document is straightforward. You need to have both your word processor and Minitab running. The document into which you wish to insert Minitab output should also be open.

The instructions in this subsection work for Microsoft Word and many other word processors.

Inserting graphical output

You can copy the graphical display in a Graph window and insert it into a word-processor document as follows.

- Click on the graph in a Session window to make it active.
- Choose **Edit > Copy** (or press **Ctrl+C**, or click the right mouse button in the Graph window and choose **Copy Graph** from the menu that is displayed). This copies the graphical display to the Windows clipboard.
- Switch to your word processor by, for example, using the task bar to switch between applications.
- Place the cursor at the point in your document where you wish to insert the graphical display.
- Finally, choose **Edit > Paste** in your word processor (or press **Ctrl+V**) and the display will be inserted into your document.

Inserting text from Minitab

You can copy text from any of the Minitab windows and insert it into a word-processor document as follows.

- Highlight the text that you want to copy.
- Choose **Edit > Copy** (or press **Ctrl+C**, or click the right mouse button in the window and choose **Copy** from the menu that is displayed). This copies the highlighted text to the Windows clipboard.
- Switch to your word processor.
- Place the cursor at the point in your document where you wish to insert the text.
- Choose **Edit > Paste** (or press **Ctrl+V**) and the text will be inserted into your document.

2 Plotting continuous variables

This chapter is associated with Subsection 3.4 of Unit 1.

In Chapter 1, you used Minitab to obtain bar charts of data. In this chapter, you will learn how to obtain plots of continuous variables: frequency histograms in Subsection 2.1 and boxplots in Subsection 2.2.

2.1 Frequency histograms

Histograms are produced using **Graph > Histogram....** In this subsection, you will obtain histograms similar to those in Figures 4, 5 and 6 of Unit 1, which represent the percentages of adults who are members of sports clubs in 49 areas of England.

Computer activity 8 Sports club membership

The percentages of adults who are members of sports clubs in the 49 sport partnership areas of England are in the Minitab worksheet **membership.mwx**. Open this worksheet now.

- Select **Graph > Histogram...**. The **Histograms** dialogue box will open.
- Select **Simple** (the default option), then click on **OK** to obtain the **Histogram: Simple** dialogue box.
- Enter **Percentage** in the **Graph variables** field at the top of the dialogue box.

The default options produce a histogram with the bars drawn vertically, the heights of the bars representing the frequencies in the intervals, ticks on the horizontal axis at the midpoints of the intervals, and the intervals chosen automatically by Minitab.

- Click on **OK**, and a histogram will be produced. Delete its title. It is shown in Figure 6 and is essentially Figure 6 of Unit 1.

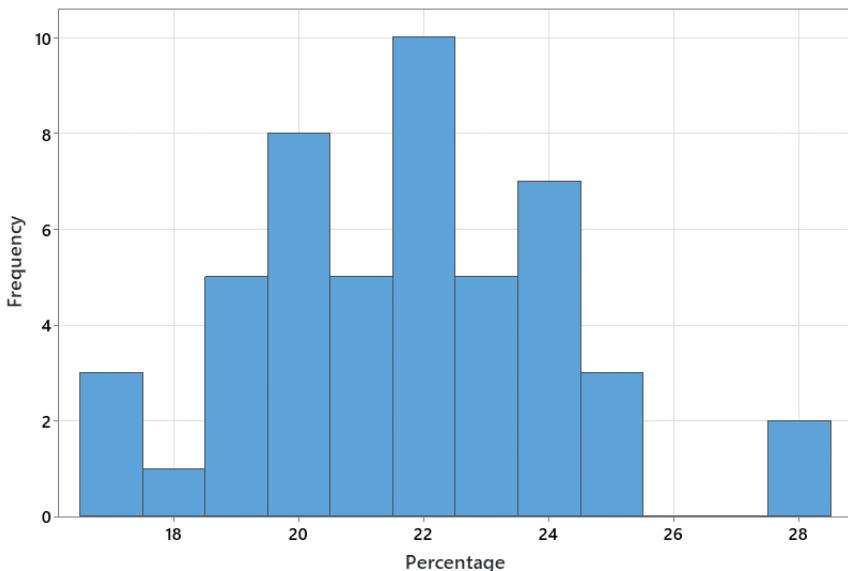


Figure 6 Default Minitab histogram of the sports club membership data

Using the default options does not always produce a histogram with intervals that you would have chosen yourself. You can specify the intervals to be used by editing the bars on this histogram, as follows.

- Double-click on the histogram to bring up the **Edit Graph** panel.
- Click on the bars to select them, then double-click on them (or press **Ctrl+T**).
- In the **Edit Bars** dialogue box, click on the **Binning** tab to view the **Binning** panel.

To produce a histogram similar to Figure 4 of Unit 1, you first need to say whether you would like ticks at the midpoints or at the cutpoints of the intervals.

- Under **Interval Type**, select **Cutpoint** to produce ticks at the borderlines between intervals.

Next you must specify the positions of the cutpoints.

- Under **Interval Definition**, select **Midpoint/Cutpoint positions**.

In the **Midpoint/Cutpoint positions** field, you must enter the positions of the cutpoints. In the histogram in Figure 4 of Unit 1, the first interval starts at 16, the last interval ends at 28, and each interval is length 1. You can set the intervals to be the same in Minitab as follows.

- Type **16:28/1** in the field. That is, enter: first cutpoint, colon, last cutpoint, forward slash, interval width.
- Click on **OK**, and the histogram in Figure 7 will be produced.

You could type 16 17 18 19 20 21 22 23 24 25 26 27 28 in the field, but this approach is usually too tedious.

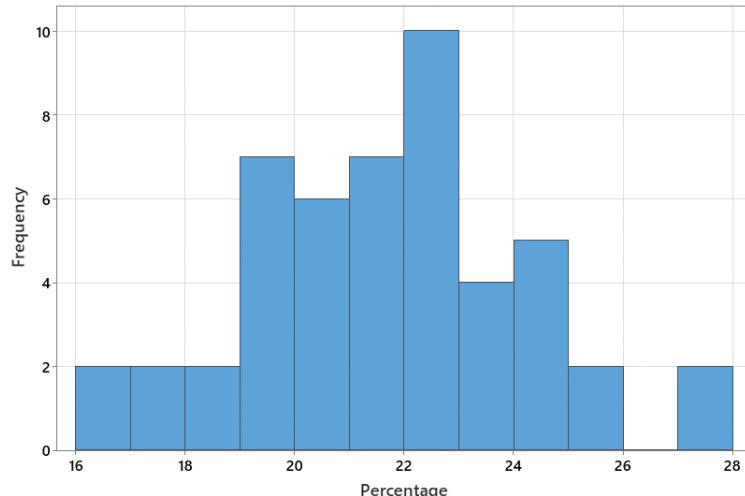


Figure 7 Another histogram of the sports club membership data

Computer activity 9 Using different intervals

Obtain a histogram for the sports club membership data similar to Figure 5(c) of Unit 1, that is, with intervals of width 2 percentage points and with the first cutpoint at 15 and the last cutpoint at 29.

Graph > Histogram...

Computer activity 10 Wages of production-line workers

The Minitab worksheet **uswages.mwx** contains data on the annual wages (in multiples of US\$100) of a random sample of 30 production-line workers in a large American industrial firm around 1980.

- Obtain a frequency histogram of the data with first cutpoint at 100 and last cutpoint at 160, and using intervals of width 5.
- Comment on the distribution of wages among the workers.

Many of the options available when using the **Histogram: Simple** dialogue box are similar to those available when using **Bar Chart....**

2.2 Boxplots

In Figure 13 of Unit 1, a boxplot of the sports club membership data was given. These data will be used in Computer activities 11 and 12 to describe the use of Minitab to produce a boxplot.

Computer activity 11 Membership of sports clubs: producing a default boxplot

The data on percentages of adults who are members of sports clubs in different English areas are in the Minitab worksheet **membership.mwx**. Open this worksheet now, or make it the active worksheet if it is already open. **In this activity you will obtain Minitab's default version of a boxplot for these data.**

- Select **Graph > Boxplot....** The **Boxplots** dialogue box will open.
- Under **One Y**, select **Simple** (the default option) and click on **OK**.

The **Boxplot: One Y, Simple** dialogue box will open. Essentially the fields and buttons in this dialogue box work in much the same way as they do when producing a histogram. (Further options for customising a boxplot are available once it has been produced.) Continue on to obtain a default boxplot for the sports club membership data as follows.

- Enter **Percentage** in the **Graph variables** field at the top of the dialogue box.

- Click on **OK**.
- Finally, delete the title, and the boxplot in Figure 8 will be produced.

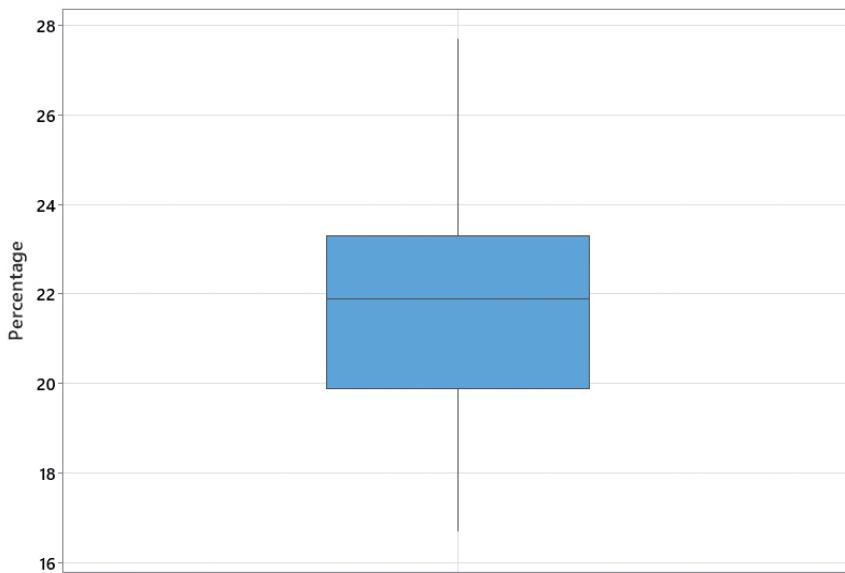


Figure 8 Default Minitab boxplot for sports club membership data

Compare this boxplot with the boxplot for this dataset in Figure 13 of Unit 1.

The vertically drawn boxplot is produced by Minitab as a default; in other words, the program always draws boxplots vertically unless you ask it to draw them horizontally. There is no hard-and-fast rule about whether boxplots should be drawn vertically or horizontally: it is largely a question of personal preference. But in M248, it has been decided to use horizontal boxplots.

Computer activity 12 Sports club membership: producing a horizontal boxplot

To obtain a horizontal boxplot instead of a vertical boxplot, you need to change one of the settings from the default.

- Obtain the **Boxplots** dialogue box.
- Select **Simple** under **One Y** and click on **OK**.
- In the **Boxplot: One Y, Simple** dialogue box, click on **Scale...** to open the **Boxplot: Scale** dialogue box.

Graph > Boxplot...

- To draw a horizontal boxplot, the option **Transpose value and category scales** in the **Axes and Ticks** panel of the **Boxplot: Scale** dialogue box must be selected. Select it now by clicking on it (or on its tick box).
- Click on **OK** to close the dialogue box, then click on **OK** in the **Boxplot: One Y, Simple** dialogue box to produce the boxplot. It is shown in Figure 9, with the title deleted.

This boxplot is the same as the one in Figure 13 of Unit 1.

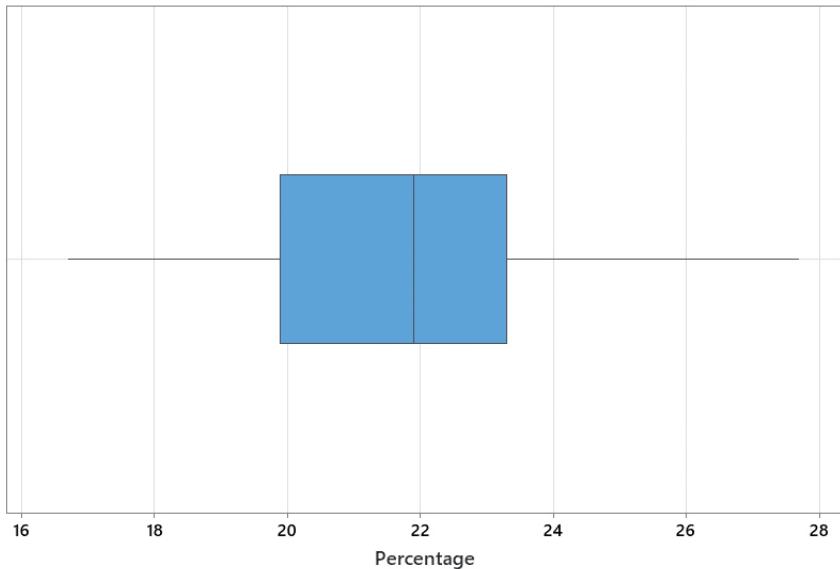


Figure 9 Horizontal boxplot for sports club membership data

Alternatively, you can produce a horizontal boxplot by editing the vertical boxplot that you obtained in Computer activity 11, as follows.

- Double-click on the boxplot to bring up the **Edit Graph** panel.
- Select the vertical axis (or 'Y Scale') on the boxplot, then double-click on it (or press **Ctrl+T**) to open the **Edit Scale** dialogue box.
- In the **Scale** panel of the **Edit Scale** dialogue box, select **Transpose value and category scales**.
- Click on **OK** and a horizontal boxplot will be produced.

3 Numerical summaries

This chapter is associated with Subsection 4.5 of Unit 1.

In Minitab, numerical summaries are produced using **Basic Statistics** from the **Stat** menu. You can either produce a fixed list of numerical summaries, or select the summaries you want from a list. The results can then be either displayed in the Session window or stored in the worksheet. Some of the datasets discussed in Unit 1 will be used to illustrate the use of Minitab to find summary statistics.

Computer activity 13 Family sizes of Ontario mothers

The data on the family sizes of particular Ontario mothers in 1941 are contained in the Minitab worksheet **family-size.mwx**. Open the worksheet now. The data for the mothers with seven years or more of education are in a column named **Long**; these are the data introduced in Example 3 of Unit 1, for which a bar chart was provided in Example 8 of the unit.

Follow the instructions below for using **Basic Statistics** to find numerical summaries of these data.

- Select **Stat > Basic Statistics > Display Descriptive Statistics....**
- Enter **Long** in the **Variables** field of the **Display Descriptive Statistics** dialogue box and click on **OK**.

The following output will be displayed in the Session window.

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Long	35	0	4.800	0.668	3.954	0.000	2.000	4.000	6.000	16.000

Some of the statistics displayed are self-explanatory. The sample mean and standard deviation are 4.8 and 3.954, respectively, and the maximum and minimum values are 16 and 0, respectively.

Minitab uses **Q1** and **Q3** as its notation for the sample lower quartile q_L and the sample upper quartile q_U , respectively. In this instance, the sample lower quartile is 2, and the sample upper quartile is 6, so the sample interquartile range is $6 - 2 = 4$.

SE Mean is the *standard error of the mean*. Don't worry if you have not met this quantity in your previous study of statistics. It is a quantity that you will meet later in M248; for the present you only need to note that it is given, but not worry about what it is. **N** is the number of values in the column, and **N*** is the number of missing values.

The data in the column **Short** are not needed in this chapter.

Stat > Basic Statistics >
Display Descriptive
Statistics...

You can specify which summary statistics are displayed by selecting from a list, as follows.

- Obtain the **Display Descriptive Statistics** dialogue box.
- Click on the **Statistics...** button to open the **Display Descriptive Statistics**: **Statistics** dialogue box. This dialogue box contains a list of the numerical summaries that can be displayed and a tick box for each summary.
- Deselect **SE of mean** and **N missing** (by clicking on them or on their tick boxes).
- Instead of working out the sample interquartile range from the values of the lower and upper quartiles for yourself, you can get Minitab to do it for you by selecting **Interquartile range** (by clicking on it or its tick box).
- Click on **OK** to close this dialogue box, then click on **OK** again.

The following output is now displayed in the Session window.

Statistics

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
Long	35	4.800	3.954	0.000	2.000	4.000	6.000	16.000	4.000



Someone who has lost a lot of weight ... but not in just two weeks!

Try this out now if you wish.

Notice that **N*** and **SE mean** are not included in the output but the sample interquartile range, which Minitab calls **IQR**, is. This set of outputs will continue to be the result of using **Display Descriptive Statistics...** until you either change the statistics selected again or start a new Minitab project or session.

Most of the numerical summaries that are available when **Display Descriptive Statistics...** is used can also be obtained using **Store Descriptive Statistics...** from the **Basic Statistics** submenu. The numerical summaries that can be displayed are available via the **Statistics...** button in the **Store Descriptive Statistics** dialogue box.

Using **Store Descriptive Statistics...** results in the output being stored in the first available columns of the worksheet instead of being collected together in the Session window.

Computer activity 14 Weight change in a clinical trial

When using **Display Descriptive Statistics...** you can obtain a graphical display of the data at the same time as numerical summaries. In Example 5 of Unit 1, data from response inhibition training for a weight-loss clinical trial were described. The data for all 83 participants in the trial are given in the Minitab worksheet **response-inhibition.mwx**. Open this worksheet now. The data on the weight changes of participants after the first two weeks of the trial are given in the column **Weight change**. The missing values are coded *****.

- (a) Follow the instructions below for obtaining a horizontal boxplot of weight change for all the participants, together with the following numerical summaries of the data: number of values, number of missing values, sample mean, sample median, sample standard deviation and sample interquartile range.

- Obtain the **Display Descriptive Statistics** dialogue box.
- Enter **Weight change** in the **Variables** field.
- Now click on the **Graphs...** button to see what displays are available.
- Select **Boxplot of data** from the list and click on **OK**.
- Click on the **Statistics...** button and in the resulting dialogue box select the required options. Click on **OK**.
- Click on **OK** in the **Display Descriptive Statistics** dialogue box.

Stat > Basic Statistics > Display Descriptive Statistics...

The numerical summaries will be displayed in the Session window and a vertical boxplot will be displayed in a Graph window. Edit the boxplot so that it is displayed horizontally (see the end of Computer activity 12 for instructions on how to do this).

- (b) Using the boxplot, comment on the shape of the weight change data.
 (c) Using suitable numerical summaries, comment on the typical weight change and the spread of weight changes experienced by all participants in the first two weeks of the clinical trial.

Computer activity 15 More on obtaining numerical summaries

In Computer activity 14, you considered the weight changes in a clinical trial for all the participants. However, of much greater importance is whether the weight changes of people in the treatment and control groups were different. One way of beginning to investigate this is to calculate numerical summaries for the treatment and control groups. In Minitab this could be done by separating the data for the treatment and control groups into different columns on the worksheet (or indeed different worksheets).

However, there is a better way, which you will use in this activity.

Whether each participant in the trial was in the treatment or the control group is given in the column **Group**: a value of 1 denotes the treatment group, and a value of 0 denotes the control group. The data for these two groups can be separated in Minitab as follows.

- Make sure the Minitab worksheet **response-inhibition.mwx** is open in Minitab.
- Obtain the **Display Descriptive Statistics** dialogue box.
- Enter **Weight change** in the **Variables** field.
- Enter **Group** in the **By variables (optional)** field.

Stat > Basic Statistics > Display Descriptive Statistics...

- Click on the **Statistics...** button and in the resulting dialogue box make sure the following options are selected: **Mean**, **Standard deviation**, **Median**, **Interquartile range**, **N nonmissing**, **N missing**.
- Click on **OK** to return to the **Display Descriptive Statistics** dialogue box.
- As boxplots are not required for this activity, click on the **Graphs...** button and make sure none of the options are selected.
- Click on **OK** to close the dialogue box, and then click on **OK** again to produce the results.

In the Session window, the numerical summaries for each group are given in separate rows. Looking at these summaries, does there appear to be a difference between the two groups?

4 Comparing variables graphically

This chapter is associated with Subsection 5.5 of Unit 1.

In this chapter, you will produce plots that will help you to compare variables graphically, as discussed in Section 5 of Unit 1. Specifically, side-by-side bar charts are discussed in Subsection 4.1, unit-area histograms in Subsection 4.2, comparative boxplots in Subsection 4.3 and scatterplots in Subsection 4.4.

4.1 Side-by-side bar charts

Side-by-side bar charts are produced in Minitab using an option within the **Bar charts** dialogue box, as described in the next activity.

Computer activity 16 Quality of tattoo removal for different depths and methods

In Computer activity 4, you used Minitab to display the results on the quality of tattoo removal – these are scores taking values 1–4, ‘1’ representing poor removal, ‘4’ representing excellent removal – without taking any of the other variables in the Minitab worksheet **tattoos.mwx** into account. Now suppose that you wish to display the results on the quality of tattoo removal separately for deep tattoos and for tattoos of moderate depth, and so compare the success of the surgical procedure for tattoos of different depths. Figure 10 represents the same data on quality of tattoo removal as Figure 3, but with the results for deep tattoos and for tattoos of moderate depth represented separately using bars drawn side by side on the same diagram.

This side-by-side bar chart shows that the results of the surgical procedure were generally better for tattoos of moderate depth than for deep tattoos.

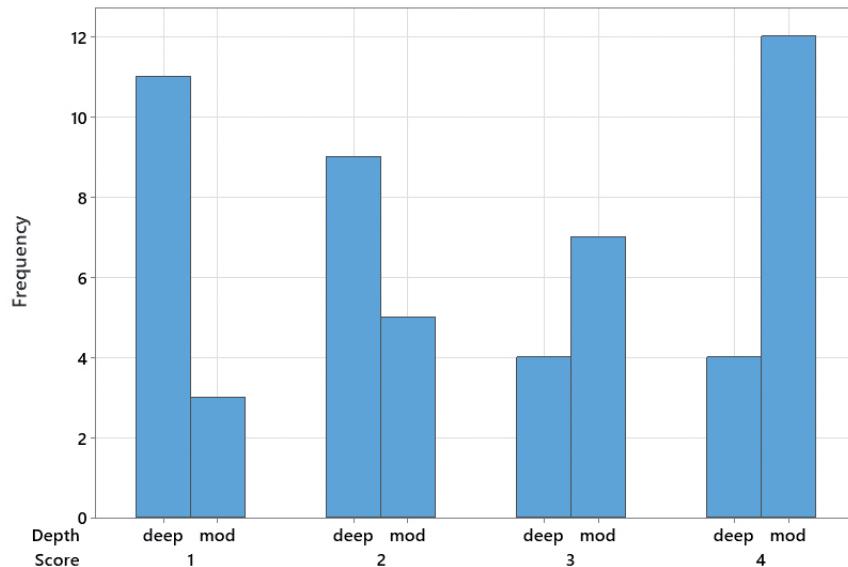


Figure 10 Quality of removal for tattoos of different depths

To obtain a side-by-side bar chart like the one in Figure 10, with the results for different groups represented by adjacent bars on the same diagram, you need to select the **Cluster** option in the **Bar Charts** dialogue box. Open the Minitab worksheet **tattoos.mwx** and follow the instructions below to obtain this bar chart for yourself.

- Obtain the **Bar Charts** dialogue box.
- Select **Counts of unique values** and **Cluster** (by clicking on its diagram).
- Click on **OK** and the **Bar Chart: Counts of unique values, Cluster** dialogue box will open.

Next you must specify the variables to be used in the **Categorical variables (2-4, outermost first)** field. You must specify the variable to be displayed first (that is, **Score**). You wish to display the results separately for tattoos of different depths, so next you must specify the variable that contains the depths (that is, **Depth**).

- Enter **Score Depth** in the **Categorical variables (2-4, outermost first)** field. (You must enter the variables in that order – the order is important.)
- Click on **OK**, and a side-by-side bar chart will be displayed in a Graph window.

To reproduce Figure 10 exactly, you need to delete the title and edit the label on the vertical axis, as described in Computer activity 5.

Now produce a side-by-side bar chart to display separately the results on

Minitab uses the word ‘cluster’ to refer to side-by-side bar charts.

Graph > Bar Chart...

quality of tattoo removal for the two surgical methods (rather than the depths of the tattoos). Comment on what the chart you obtain tells you.

4.2 Unit-area histograms

In Subsection 2.1, you saw how to use Minitab to produce frequency histograms. As you will find in this subsection, unit-area histograms can be produced in a very similar way.

Computer activity 17 A unit-area histogram of sports club membership

In Computer activity 8, you created a frequency histogram of the sports club membership data (Example 4 in Unit 1). In this activity, you will create the corresponding unit-area histogram.

- Open the Minitab worksheet **membership.mwx** in Minitab.
- Obtain the **Histogram: Simple** dialogue box.
- Enter **Percentage** in the **Graph variables** field.

Graph > Histogram... and select **Simple**.

So far, this is the same as you would do to create a frequency histogram. To obtain a unit-area histogram, the following needs to be done.

- Click on the **Scale...** button to obtain the **Histogram: Scale** dialogue box.
- On the **Y-Scale Type** tab, the **Frequency** option is selected. This means that, as things stand, Minitab would produce a frequency histogram. To switch to producing a unit-area histogram, select **Density** instead.
- Click on **OK** to close the dialogue box and then click on **OK** again, and a density histogram will be displayed in a Graph window.

As with frequency histograms, the intervals displayed will be ones chosen by Minitab. These can be overridden in exactly the same way as for frequency histograms. For example, use the same intervals as in Figure 23 in Unit 1 by doing the following.

- Double-click on the histogram to bring up the **Edit Graph** panel.
- Obtain the **Edit Bars** dialogue box. (Click on the bars to select them and then double-click or press **Ctrl+T**.)
- On the **Binning** tab, select **Cutpoint** and **Midpoint/Cutpoint positions**. Then enter **16:28/1** in the **Midpoint/Cutpoint positions** field.
- Click on **OK**.

The resulting histogram, shown in Figure 11, matches Figure 23 in Unit 1. (This unit-area histogram is the same as the frequency histogram in Figure 7 except for the vertical rescaling.)

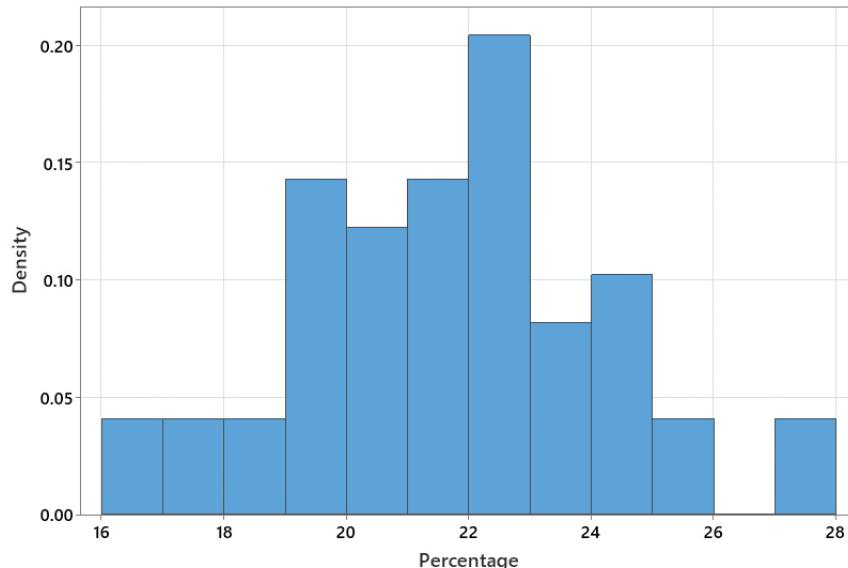


Figure 11 Unit-area histogram of sports club membership data

Computer activity 18 A unit-area histogram of weight change

In Computer activity 14, you created a boxplot of the weight change data from a clinical trial alongside some numerical summaries. These data are given in the Minitab worksheet **response-inhibition.mwx**.

- Create a unit-area histogram of weight change for all the participants in the clinical trial. You should use the following cutpoints: $-8, -7, -6, \dots, 3, 4$.
- Using the histogram, comment on the shape of the data. In particular, do the data appear to be unimodal, bimodal or multimodal?

4.3 Comparative boxplots

Producing comparative boxplots in Minitab is an extension of producing single boxplots. The option selected in the **Boxplots** dialogue box depends on how the data are arranged in the Minitab worksheet.

Computer activity 19 Comparative boxplot of the weight change data

Recall that in Computer activity 14, you produced a boxplot of weight change in two weeks for all the participants in a weight-loss clinical trial, and that in Computer activity 15, you calculated numerical summaries for these data for the treatment and control groups separately. In this activity, you are going to produce a comparative boxplot for these same data.

- (a) Open the Minitab worksheet **response-inhibition.mwx**. How are the weight change data for treatment and control groups structured in Minitab?
- (b) Obtain (horizontal) comparative boxplots for the data by doing the following. Note that these instructions will produce a boxplot of the data in the column **Weight change** for each distinct value in the column **Group**. Here there are two different values, 0 and 1, in the column **Group**, so two boxplots will be produced on a single diagram.

Graph > Boxplot...

- Obtain the **Boxplots** dialogue box.
- The data are in a single column, with groups given in a second column, so to obtain a comparative boxplot, select **With Groups** under **One Y** and click on **OK**. The **Boxplot: One Y, With Groups** dialogue box will open.
- Enter **Weight change** in the **Graph variables** field and **Group** in the **Categorical variables for grouping (1-4, outermost first)** field.
- For horizontal boxplots, click **Scale...**, and in the **Boxplot: Scale** dialogue box, select **Transpose value and category scales** in the **Axes and Ticks** panel.
- Click on **OK** to close this dialogue box, then click on **OK** again to produce the comparative boxplot.

Next, give the two boxplots more meaningful labels on the vertical axis, as follows.

- Double-click on the boxplot to bring up the **Edit Graph** panel.
- Click on the vertical axis to select the vertical scale, and press **Ctrl+T** (or double-click on it while it is selected). (This is still called the ‘X Scale’ even though you have transposed Minitab’s default axes.)
- Click on the **Labels** tab in the **Edit Scale** dialogue box to bring the **Labels** panel uppermost.
- Under **Major Tick Labels**, select **Specified** and type **Control Treatment** in its field. (These must be input in that order.)
- Click on **OK** and the labels will change.
- (c) Use the comparative boxplot to compare the weight changes in two weeks of the clinical trial for participants in the treatment and control groups.

For the weight change data in Computer activity 19, all the numerical data to be represented by boxplots were in one column. A variable in a second column indicated the different groups that were to be plotted separately. When the data are in this format, the **With Groups** option under **One Y** must be selected in the **Boxplots** dialogue box. In the next activity, you will produce a comparative boxplot when the data are in a different format.

Computer activity 20 Comparative boxplot of memory recall data

The memory recall times used in Activities 23 and 24 of Unit 1 are in the Minitab worksheet **memory.mwx**. Open this worksheet now.

In this worksheet, the two batches of data are in separate columns called **Pleasant** and **Unpleasant** (indicating what type of memory the subjects were trying to bring back). To obtain a comparative boxplot for data arranged like this, a different option in the **Boxplots** dialogue box is used.

- Obtain the **Boxplots** dialogue box.
- Under **Multiple Y's**, select **Simple**, then click on **OK** to open the **Boxplot: Multiple Y's, Simple** dialogue box.

The name of each column for which a boxplot is required must be entered in the **Graph variables** field. The boxplot for the variable entered first will appear at the top of a horizontal comparative boxplot; obtain such a comparative boxplot with the boxplot for the pleasant memories at the top, as follows.

- Enter **Pleasant Unpleasant** in the **Graph variables** field.
- Click on **Scale...** and select **Transpose value and category scales** in the **Axes and Ticks** panel of the **Boxplot: Scale** dialogue box.
- Click on **OK** to close this dialogue box, then click on **OK** again to produce the comparative boxplot.

Finally, delete the title and change the label on the horizontal axis to **Recall time (seconds)**. Check that the resulting plot looks similar to Figure 27 of Unit 1.

You will be using these data to produce boxplots like those in Figure 27 in Activity 24 of Unit 1.

Graph > Boxplot...

Computer activity 21 More boxplots of the weight change data

In Computer activity 19, you produced a comparative boxplot of weight change in the treatment and control groups in the first two weeks of a clinical trial. In this activity, you will produce a comparative boxplot that also displays information on whether the weight change was different for the male and female participants. To do this, more than one categorical variable can be entered in the **Categorical variables for grouping (1-4, outermost first)** field in a similar way as you did in the **Categorical variables (2-4, outermost first)** field in Computer activity 16.

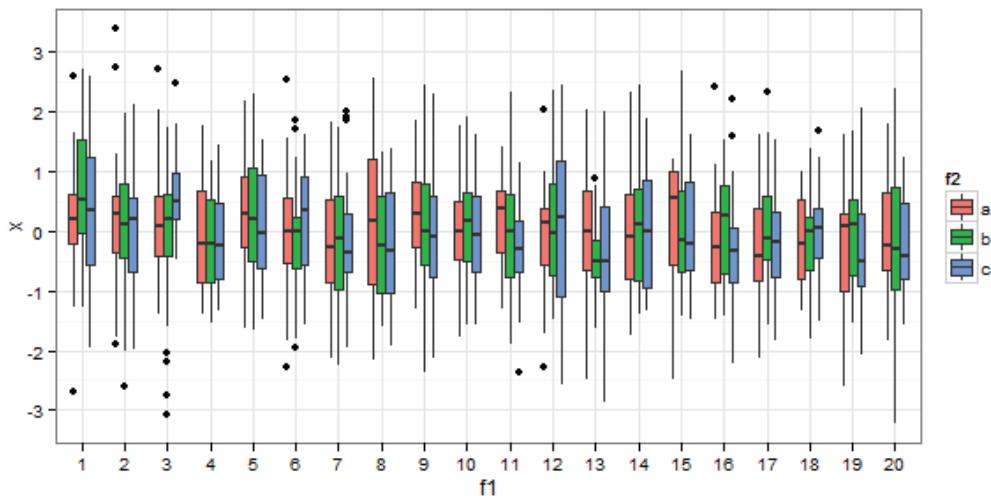
Open the Minitab worksheet **response-inhibition.mwx** now if it is not already open.

- (a) Produce a (horizontal) comparative boxplot for weight change in the treatment and control groups for females and males separately. Edit the labels for the individual boxplots so that the new labels for group

are **Treatment** and **Control**, and the new labels for gender are **Female** and **Male**. (The treatment groups are 0 = Control and 1 = Treatment; the gender groups are 1 = Female, 2 = Male.)

- (b) Does there appear to be a difference between the weight changes experienced by the females and males in the different treatment groups in this clinical trial?

As the following figure shows, even boxplots become difficult to interpret and compare if there are too many of them!



4.4 Scatterplots

Scatterplots are produced using **Scatterplot...** from the **Graph** menu.

Computer activity 22 Relationship between field and laboratory pipeline defect depth measurements

The Minitab worksheet **alaska.mwx** contains the data underlying Example 25 of Unit 1. Open this worksheet now. The columns contain field and laboratory defect depth measurements for 107 defects in the Trans-Alaska oil pipeline.

Follow the instructions below to obtain a scatterplot similar to Figure 28 of Unit 1.

- Choose **Graph > Scatterplot....**
- In the **Scatterplots** dialogue box, select **Simple** (the default) and click on **OK**.

The **Scatterplot: Simple** dialogue box is similar to those for producing simple bar charts, histograms and boxplots. The main difference is in the area at the top: you must enter the names of the two variables to be plotted in row 1 under **Y variables** and **X variables**.

- Enter **Field defect depth** under **Y variables** and **Laboratory defect depth** under **X variables**. The quickest way to this is to double-click on **Field defect depth**, then double-click on **Laboratory defect depth**.
- Click on **OK** to produce a scatterplot. Finally, delete the title.

The scatterplot produced is shown in Figure 12, the Minitab version of Figure 28 of Unit 1.

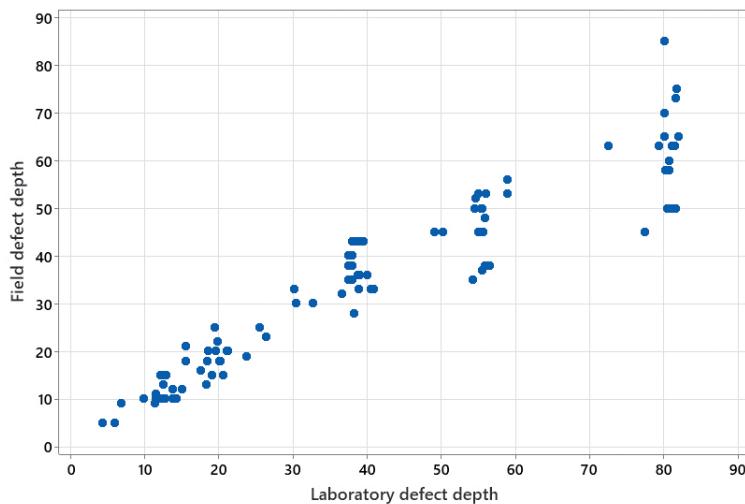


Figure 12 Field and laboratory measurements of defect depth

Computer activity 23 Comparing distances

The Minitab worksheet **distance.mwx** gives the distances between twenty pairs of locations in Sheffield. For each pair of locations, two distances are given (in miles): the distance by road (variable **Road**) and the ‘straight line’ distance given on a map (variable **Map**). Open this worksheet now.

- Produce a scatterplot of road distances against map distances. (That is, plot the road distances on the vertical axis and the map distances on the horizontal axis.) Edit the axes so that they are labelled **Distance by road** and **Distance on a map**.
- Use the scatterplot to comment on the relationship between the map and road distances. Also, do there appear to be any outliers? If so, in what way are they unusual?



‘As the crow flies’ is a long-established idiom for the shortest distance between two points. However, crows don’t fly in straight lines over long distances!



5 From samples to models for discrete data

This chapter is associated with Subsection 2.2 of Unit 2.

In this chapter, you will use one of the M248 animations to explore the settling-down phenomenon that was mentioned in Subsection 2.2 of Unit 2.

The animation is designed to allow you to explore the situation in which an unbiased six-sided die is rolled repeatedly and the relative frequencies of the outcomes of the rolls of the die are noted after each roll. Physically rolling a die and recording the outcome after each roll would quickly become very tedious. Fortunately, a computer can help with this sort of task. Of course, the computer does not actually roll a die. But when it is programmed appropriately, it produces results that are indistinguishable from what might occur if a die were rolled. This sort of alternative to carrying out a real experiment is known as **simulation**.



In football, simulation means trying to gain an unfair advantage

In a long sequence of repetitions of a study or experiment, bar charts and histograms of samples tend to settle down towards probability distributions.

In the simulation you will perform in Computer activity 24, you will explore how this works out for a situation involving discrete data. That is, you will see how bar charts of samples from a particular situation involving discrete data settle down to a probability mass function as the number of repetitions of the experiment increases.



Computer activity 24 The score on a die

The animation **Score on a die** may be used to simulate an experiment in which an unbiased six-sided die is rolled a large number of times and the observed frequencies of 1s, 2s, ..., 6s (the six possible outcomes) are recorded.

- Open the **Score on a die** animation.

The animation starts with a blank graph with the possible numbers that can be observed on an unbiased six-sided die (i.e. 1, 2, ..., 6) on the horizontal axis, and the relative frequency (from 0 to 1) on the vertical axis.

The controls for running the animation are underneath this graph. Initially, the animation is set to roll the die 30 times (**Sample size, n** is 30). Above this, there is a slider that controls how fast the die is rolled: this is initially set to **Slow**. And above that, the number of ‘rolls’ of the die is recorded (**Number of rolls**). The simulation can be stopped at any point by clicking on **Stop**, and, as might be expected, clicking on **Reset** clears the graph ready for another simulation.

The results of this activity are discussed next.

(a) Start by ‘rolling the die’ 30 times.

- Make sure the slider is at **Slow**, and click on **Start** to begin the simulation.

After each roll of the die, a bar chart of the relative frequencies of the scores 1, 2, 3, 4, 5 and 6 is displayed on the graph. This is the equivalent of a probability mass function. When you are sure that you understand what is happening, speed up the simulation by moving the slider towards **Fast**.

- Click on **Reset**, then **Start** to run another simulation.
- Run several simulations with 30 rolls of the die.

What do you notice about the relative frequencies in your simulations?

(b) Now change the number of rolls (**Sample size, n**) to 300 and run a simulation. Then run a simulation for 1000 rolls of the die.

What do you notice about the relative frequencies of the scores as the total number of rolls increases?

Figure 13 shows the relative frequencies produced by the animation after each of 30, 300 and 1000 simulated rolls of the die, as in Computer activity 24.

The relative frequencies that you produce are likely to look different to these because they are based on simulated random rolls.

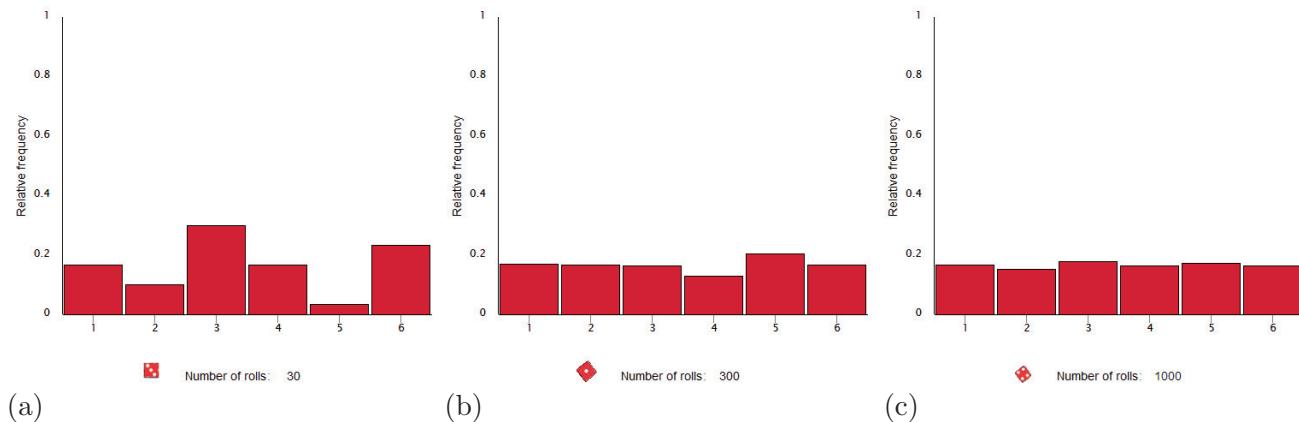


Figure 13 Relative frequencies from a simulation of rolling an unbiased die: (a) 30 rolls, (b) 300 rolls, (c) 1000 rolls

It is apparent from the diagrams in Figure 13 that the computer model is capable of reflecting the random variation in the physical process. Although, for example, we ‘expect’ 50 observations of each of the six equally likely outcomes after 300 rolls of the die, which would mean that the relative frequencies were exactly the same for each outcome, we would nevertheless be somewhat surprised to see exactly 50 observations of each outcome. In the simulation, the relative frequencies are not exactly the same for each outcome, even after 1000 ‘rolls of the die’.

If we were to carry out a very extended experiment involving many rolls of a die, then, assuming that the die is unbiased and applying a symmetry argument, we would expect to obtain each of the outcomes 1, 2, 3, 4, 5, 6 an equal proportion of times as discussed in Example 11 of Unit 2. That is, the diagrams in Figure 13 are becoming more and more like that in Figure 14 as the number of rolls of the die increases.

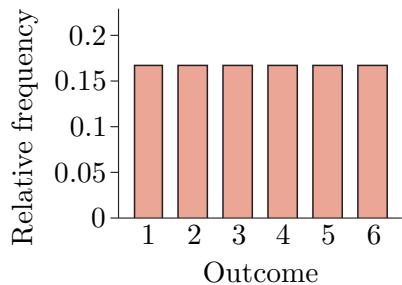


Figure 14 Relative frequencies of outcomes of an unbiased die rolled an infinite number of times



6 From samples to models for continuous data

This chapter is associated with Subsection 2.3 of Unit 2.

In Chapter 5, you used the animation **Score on a die** to explore the settling-down phenomenon in a discrete case. A similar effect occurs in the continuous case; in this chapter you will use the animation **American weights** to explore this in the context of a dataset concerning people's weights. You will see how, for very large samples, the shape of a histogram varies very little from sample to sample. So, given a very large sample, the relative frequencies of the different possible outcomes may be used to estimate the probabilities of the outcomes, thus suggesting a form for the probability distribution for the random variable.



Computer activity 25 The weights of American adults

The weights of American adults vary from one individual to another, so the weight of a randomly selected American adult may be modelled by a random variable.

Although in practice weights can be recorded only as precisely as the scales used, weights can take any value in an interval of values, so a continuous model is needed. In this activity, you will investigate what a continuous model for the weights of American adults might look like.

The data on which the animation **American weights** is based are from the third National Health and Nutrition Examination Survey (NHANES III). This survey was carried out in the period 1988–94. The animation may be used to simulate taking samples of weights (in kilograms) of American adults from a population of such weights.

- Open the **American weights** animation.

The animation allows you to take samples from the population of weights and display histograms of the sample data. The histograms are all unit-area histograms, scaled so that the total area of the bars is equal to 1.

The animation starts with a single blank graph ready to plot the first unit-area histogram. The horizontal axis is weight (in kg), but is unlabelled to avoid overcrowding as more histograms are added to the screen.

The controls for running the animation are underneath this graph.

Initially, the sample size is set to be 500 (**Sample size, n** is 500). There are two other buttons: **Take sample**, which you need to click in order to take a sample and plot a histogram, and **Reset**, which clears any histograms and returns to the initial blank graph.

The results of this activity are discussed below.

- Take four samples of size 500. Comment on the shapes of the four histograms. (Note that the vertical scale is not always the same on each figure within each tab.)
- Change the sample size to 5000 and take four samples of size 5000. Changing the sample size to 5000 creates a new tabbed panel labelled **$n=5000$** for these histograms; the original histograms for $n = 500$ remain on the tabbed panel labelled **$n=500$** .
Comment on the shapes of the histograms for sample size 5000.
Notice that since the sample size is much larger than that in part (a), narrower intervals have been used when drawing the histograms, thus providing more information about the distribution of weights in the population.
- Change the sample size to 50 000 and take four samples of size 50 000. Once again, comment on the shapes of the histograms. Since the sample size is much larger than that in part (b), even narrower intervals have been used when drawing the histograms, thus providing yet more information about the distribution of weights in the population.
- What do you think a model for the weights of American adults should look like?

More recent data suggest a shift upwards in these weights of the order of 5 kg.

Figure 15 shows histograms produced by the animation after taking samples of size 500, 5000 and 50 000, as in Computer activity 25.

For each of the sample sizes examined, all the histograms have a single clear peak and a longer tail of values to the right of this peak than to the

A new tabbed panel labelled **$n=50000$** will be added to the animation.

The histograms that you produce are likely to look different to these because they are based on simulated random samples.

left. So a model for the weights of American adults should be unimodal and right-skew. For the larger sample sizes there was less variation between the shapes of the histograms; and since more intervals were used to summarise the data, more detailed information about the distribution of weights in the population could be deduced from a histogram. A possible model for the weights of American adults is explored in Computer activity 26.

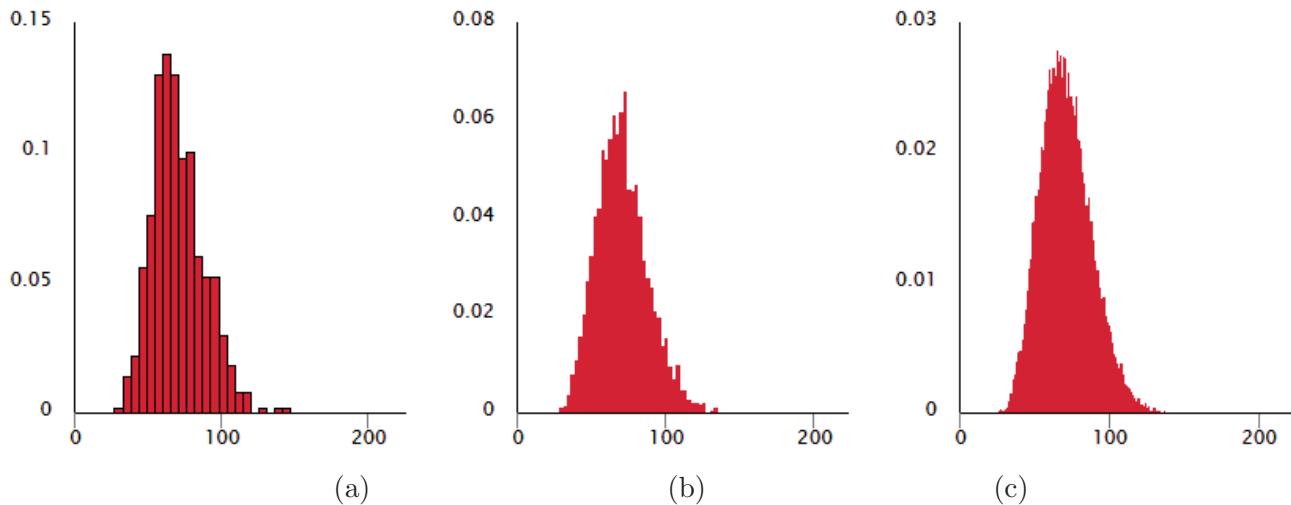


Figure 15 Histograms from samples of simulated weights of size: (a) 500, (b) 5000, (c) 50 000



Computer activity 26 Modelling weights

You should still have the animation **American weights** running. If not, then start it now.

- Click on the tab labelled **Settling down**.

This **Settling down** panel works in the following way: each time a sample is taken, the results are combined with those of samples already taken while using the panel, so that the results of successive samples are accumulated to produce one large histogram.

- Begin by taking a sample of size 500, then take several more samples of this size.
- Increase the sample size (to 5000, say) and take several more samples.

Notice how the shape of the histogram changes as each sample is added: it becomes less jagged and changes less with each additional sample as the total number of weights sampled increases.

- Increase the sample size again (to 10 000, say) and keep taking samples.

Notice that after a while the shape of the histogram changes very little as further results are accumulated: the shape of the histogram is settling down and most of the jaggedness, which occurred when smaller numbers of weights were sampled, disappears. This suggests that a smooth curve

might provide an adequate model for the variation in the weights of American adults.

- Click on the **Add model** tick box and a curve will be superimposed on your histogram.

This curve ‘fits’ the data well: the curve matches the shape of the histogram very closely. The total area under this curve is equal to 1. This means that, for example, the proportion of American adults weighing over 100 kg could be estimated by finding the area under the curve to the right of 100.

In the activities in this chapter and the previous one, you have been simulating situations involving chance. For example, you used the **Score on a die** animation to simulate rolling a fair six-sided die; and you used the **American weights** simulation to simulate taking random samples from populations.

It may seem paradoxical to use a computer to produce ‘random’ values: we would expect any computer program to produce output that is entirely predictable. Nevertheless, computer ‘random number generators’ are in common use; these generate sequences of ‘random’ integers. Given an initial value – the *seed* value – the sequence generated is predictable and therefore is not truly random. Numbers generated in this way are called *pseudo-random numbers*. However, in practice, sequences of such numbers are indistinguishable from sequences of random numbers, so they may be regarded as sequences of random numbers and used to simulate random samples for statistical simulations. If the sequence of numbers generated on any occasion is to be unpredictable, then it is important that an element of chance is involved in the choice of the seed value for the sequence. In each of these animations, the seed number is taken from the computer’s clock when you start the animation. This makes it extremely unlikely that the samples obtained on two different occasions will be the same.

The idea of using a curve to model the variation in a continuous random variable is developed further in the rest of Unit 2 and in later units of M248.



Random seeds?

Minitab also uses the computer’s clock to choose the seed number for generating ‘random’ numbers.

7 The binomial distribution

This chapter is associated with Subsection 2.2 of Unit 3.

In Minitab, values of the probability function and the cumulative distribution function may be obtained for a number of families of distributions. This is done using **Probability Distributions** from the **Calc** menu. The use of Minitab to find binomial probabilities will be illustrated using an example discussed in Unit 3.

Computer activity 27 Multiple choice examination scores

An examination consists of twenty multiple choice questions. For each question, the correct answer is one of five options. The random variable T ,

This is the situation discussed in Example 9 of Unit 3.

which denotes the number of correct answers obtained by a student who guesses answers at random, has a binomial distribution $T \sim B(20, 0.2)$.

In this activity you will use Minitab to find the probability that the student scores exactly 10 and so just passes the examination. Run Minitab now if it is not already running. You do not need any particular Minitab worksheet for this activity.

- Choose **Calc > Probability Distributions** so that the contents of this submenu are displayed.

Three groups of distributions are listed in the submenu. The first group contains the names of several commonly used continuous probability models; you will meet most of these in M248. The third group also consists of continuous probability models (most of which you won't meet in M248). The second group contains discrete probability models; the binomial distribution is the first one listed.

- Choose **Binomial...** from the submenu, and the **Binomial Distribution** dialogue box will open.

We wish to find the probability $P(T = 10)$, where $T \sim B(20, 0.2)$. This probability can be obtained in Minitab as follows.

- In the **Binomial Distribution** dialogue box, first select **Probability**. This tells Minitab that we wish to calculate a value of the binomial p.m.f.
- The parameters of the binomial distribution are $n = 20$ and $p = 0.2$, so type the number 20 in the **Number of trials** field, and the value 0.2 in the **Event probability** field. Note that the value of the parameter p must be entered as a decimal: Minitab does not accept fractions for parameter values.
- A single probability is required, so select **Input constant** (rather than **Input column**, which is used to calculate the p.m.f. values for a collection of values of the random variable stored in a column).
- Since the probability required is $P(T = 10)$, enter the value 10 in the **Input constant** field. You do not need to store this probability, so leave the **Optional storage** field empty.
- Click on **OK**.

The following output will be displayed in the Session window.

Binomial with n = 20 and p = 0.2

x	P(X = x)
10	0.0020314

The first thing to notice in the Minitab output concerns the output heading. In Unit 3, the term probability density function is used only when referring to the probability function of a continuous random variable; the probability function of a discrete distribution is called the probability

mass function. However, Minitab includes the word ‘Density’ in the heading in the output for both discrete and continuous distributions.

Minitab then states the distribution from which it has calculated a probability (i.e. the binomial with parameters $n = 20$ and $p = 0.2$). Minitab always refers to the random variable in a calculation as X . The final output therefore gives the value of x for which the probability is required (in this case $x = 10$) and the associated p.m.f. for that value of x (in this case $P(X = 10) = 0.0020314$). So the probability that a student who guesses answers at random will obtain exactly ten correct answers in twenty questions is approximately 0.002.

Computer activity 28 Fail the multiple choice examination?

In Computer activity 27, T , the number of correct answers obtained by a student taking a multiple choice examination who guesses answers at random, has a binomial distribution $T \sim B(20, 0.2)$. To pass the test, the student must get at least 10 out of the 20 questions correct. In this activity, you will use Minitab to obtain the probability that the student fails the examination.

If the student needs to answer at least 10 questions correctly to pass the examination, then the probability that the student fails is

$$P(T < 10) = P(T \leq 9) = F(9),$$

where $F(t) = P(T \leq t)$ is the c.d.f. of random variable T . The value of $F(9)$ can be obtained in Minitab as follows.

- Obtain the **Binomial Distribution** dialogue box. Notice that the dialogue box contains the settings from the previous calculation.
- Select **Cumulative probability**.
- The probability $P(T \leq 9)$ is required, so change the value in the **Input constant** field to 9.
- Click on **OK**.

Calc > Probability Distributions > Binomial...

You should obtain the following output in the Session window.

Binomial with $n = 20$ and $p = 0.2$

x	$P(X \leq x)$
9	0.997405

So the probability that a student who guesses answers at random will fail the examination is approximately 0.997.



Computer activity 29 Retake or resit?

Consider once again the multiple choice examination of Computer activity 27.

- Suppose that students have to retake the module the following year if they answer fewer than four questions out of twenty correctly. What is the probability that a student who guesses answers at random has to retake the module?
- Students who fail the examination (that is, score less than 10) are allowed to resit the examination without retaking the whole module if they answer at least four out of the twenty questions correctly. What is the probability that a student who guesses answers at random fails but is allowed to resit the examination?

Computer activity 30 Obtaining a table of results

For this activity, you will need a new worksheet. If you don't have one already, choose **File > New > Worksheet** to produce a new blank worksheet.

In this activity, you will obtain a table similar to Table 2 of Unit 3, containing values of the p.m.f. and the c.d.f. for the binomial distribution $B(20, 0.2)$. If you store your results in a worksheet, then values of the p.m.f. and the c.d.f. can be displayed in a single table.

To do this, you need to enter the values for which you require probabilities in a column of a worksheet *before* opening the **Binomial Distribution** dialogue box. We will enter the numbers $0, 1, 2, \dots, 20$ in column C1 of a worksheet. You can do this either by typing in the numbers directly or, as described below, using **Make Patterned Data** from the **Calc** menu.

- Choose **Calc > Make Patterned Data > Simple Set of Numbers...**
- To store the numbers in column C1, type C1 in the **Store patterned data in** field.
- All the integers from 0 to 20 are required, so enter 0 in the **From first value** field and 20 in the **To last value** field. The other fields should each contain default values of 1. (If by any chance they do not, then change the values to 1 in these fields.)
- Click on **OK** and the numbers will be stored in column C1.

Values of the p.m.f. can be entered in column C2 as follows.

- Obtain the **Binomial Distribution** dialogue box.

- Select **Probability**.
- The parameters $n = 20$ and $p = 0.2$ should still be set from the previous activity; if they are not, then enter these values.
- The column C1 stores the values for which probabilities are required, so select **Input column** and type C1 in its field.
- Now type C2 in the **Optional storage** field and click on **OK**. The results will be stored in column C2 of the worksheet.

Values of the c.d.f. can be entered in column C3 in a similar way.

- Obtain the **Binomial Distribution** dialogue box.
- Select **Cumulative probability** and type C3 in the **Optional storage** field. (There is no need to change any of the other settings.)
- Click on **OK** and the results will be stored in column C3.

Using this procedure, apart from different column headings, you should obtain the table of values shown in Table 1. The values in Table 1 match with those in Table 2 of Unit 3, which are given correct to four decimal places (except the very last entry, which is exactly 1).

Computer activity 31 will provide you with further practice at finding binomial probabilities using Minitab to do the calculations.

Table 1 The probability distribution of $B(20, 0.2)$

t	$P(T = t)$	$P(T \leq t)$
0	0.011529	0.01153
1	0.057646	0.06918
2	0.136909	0.20608
3	0.205364	0.41145
4	0.218199	0.62965
5	0.174560	0.80421
6	0.109100	0.91331
7	0.054550	0.96786
8	0.022161	0.99002
9	0.007387	0.99741
10	0.002031	0.99944
11	0.000462	0.99990
12	0.000087	0.99998
13	0.000013	1.00000
14	0.000002	1.00000
15	0.000000	1.00000
16	0.000000	1.00000
17	0.000000	1.00000
18	0.000000	1.00000
19	0.000000	1.00000
20	0.000000	1.00000

Computer activity 31 More multiple choice tests

- A test consists of ten multiple choice questions, each of which has eight options. A student must answer at least five of the questions correctly to pass the test. Find the probability that a student who guesses answers at random answers exactly five questions correctly, and so just passes the test. What is the probability that such a student fails the test?
- In another test, there are thirty questions, each of which has four options. To pass the test, a student must answer at least half of them correctly. Find the probability that a student who guesses answers at random passes the test (that is, answers at least fifteen of the questions correctly).

In this chapter, you have used Minitab to calculate probabilities involving binomial distributions. In later chapters, you will use Minitab to calculate probabilities for other probability distributions. As you will see, the procedure is essentially the same whatever the distribution.



8 Is the uniform model reasonable?

This chapter is associated with Subsection 5.1 of Unit 3.

In this chapter, you will use the animation **Royal deaths** to explore the dataset on royal deaths that was introduced in Example 20 of Unit 3. The discrete uniform distribution was proposed in the unit for the variation in the data. In this chapter, you will investigate informally whether the proposed model is a good fit to the data.



Computer activity 32 Royal deaths

In this activity, you will use an animation designed to simulate taking samples from a probability model for the month of death of descendants of Queen Victoria.

- Open the **Royal deaths** animation.

The controls for the simulation are along the bottom of the animation. Above the controls are two panels.

The bar chart at the top of the left-hand panel shows the relative frequencies of months of death (January = 1, February = 2, ..., December = 12) of 82 descendants of Queen Victoria; they all died of natural causes.

As shown in the bar chart, there were more deaths in some months than in others. Is this due to random variation and just a feature of the sample of data, or are deaths more likely to occur in some months than in others?

If a death is equally likely to occur at any time of the year, then a possible model (which ignores the unequal lengths of months) is a discrete uniform distribution on the integers 1, 2, ..., 12. A diagram of the probability mass function for this model is shown at the bottom of the left-hand panel.

Do you think that the data could be a random sample from this uniform distribution, or is there more variation in the numbers of deaths in different months than would be likely to occur by chance? You can investigate this by taking random samples from the uniform distribution and then comparing bar charts of the samples with the bar chart of the data. If you find that some of the bar charts are as jagged as that for the data, then this will mean that the model is a plausible one and that there is no reason to suppose that deaths are more likely to occur at some times of the year than at others: the jaggedness in the bar chart of the data could simply be due to random variation. Before taking any samples, make a note of what your intuition tells you: do you think that these data might be a sample from a uniform distribution?

- (a) In Chapters 5 and 6, you investigated the settling-down phenomenon: you saw that, in general, bar charts and histograms for small samples tend to be more jagged than those for large samples from the same population. As a first step towards investigating whether the uniform

distribution is a good model for the deaths data, try running the simulation for a number of different sample sizes.

- Ensure that the **Sample size, n** field is set to 50 (the default), then click on **Take sample**.

A bar chart of relative frequencies for a sample of size 50 will be displayed in a tabbed panel labelled **$n=50$** on the right-hand side of the screen.

- Take several more samples of size 50.

For each sample, a bar chart will be added to the tabbed panel labelled **$n=50$** . There is a scroll bar on the right-hand side of the animation that allows you to scroll through the bar charts.

- Now change the **Sample size, n** field to 500 and click on **Take sample**.

A new tabbed panel labelled **$n=500$** will appear that contains a bar chart of this new sample. (The bar charts for the samples of size 50 can still be seen if you click on the tab labelled **$n=50$** .)

- Take several more samples of sample size 500.
- Now increase the sample size to 5000, and take several samples.

What do you notice about how the jaggedness of the bar charts changes as the sample size is increased?

- (b) You will have seen in part (a) that for larger sample sizes the bar charts are less jagged. So if you want to compare the bar chart of the data with bar charts of samples from the model, then the size of the samples you take is important: you must take samples of the same size as the given set of data. In this case, samples of size 82 are required.

- Change the number in the **Sample size, n** field to 82 and click on **Take sample**. A bar chart of a sample of size 82 will be displayed in the right-hand tabbed panel labelled **$n=82$** .
- Click on **Take sample** several times more to obtain further samples of size 82 from the uniform distribution.

Compare the bar chart of each sample with the bar chart of the data. Do you think that a discrete uniform distribution is a good fit for the data? Or are royal deaths more likely to occur in some months of the year than in others? Does your conclusion here confirm or contradict what you thought before carrying out the simulation?

A new tabbed panel labelled **$n=5000$** will appear.

You might have been surprised by the amount of variability inherent in the discrete uniform distribution even for samples of size $n = 82$. It is this that justifies the claim in the solution to Computer activity 32 that the discrete uniform distribution is a plausible one for the distribution of the months of royal deaths. There again, as mentioned in Example 20 of Unit 3, there might be a systematic, and interpretable, deviation from uniformity over the whole year in that the data seem to suggest that the summer months



are less likely to include a death than the winter months. So there certainly remains room for doubt. This, unfortunately, is typical when dealing with random variation. We often cannot make clear-cut statements about the results of investigations. The best way to resolve the situation, if it were available to you, would be to collect more data.

Later in M248, you will meet a formal statistical method for testing the goodness of fit of a model to data.



9 The binomial and Poisson distributions

This chapter is associated with Section 1 of Unit 5.

In this chapter, you will investigate the circumstances in which a binomial distribution with parameters n and p may be approximated by a distribution that depends only on the mean $\mu = np$ of the binomial distribution – that is, by a Poisson distribution with parameter μ . This chapter begins with an example of a situation in which a binomial distribution can be approximated by a Poisson distribution.



Damage caused by a V-1 bomb in London in 1944

Example 1 V-1 flying bombs

Late in the Second World War, in the autumn of 1944, nearly 10 000 V-1 flying bombs were launched by German forces against British towns and cities. The majority of those that got through the defences descended on London. German propaganda claimed that the V-1 bomb was an accurately aimed weapon. To investigate this claim, the London Fire Brigade plotted the positions of all the V-1 hits within a 6 km by 6 km square of South London. Probability modelling was then used to investigate whether the positions where the bombs landed were randomly distributed in this square.

The 6 km by 6 km square was divided into a grid of 576 squares each of size 0.25 km by 0.25 km. The number of V-1 bombs that landed on the 576 grid squares was 537 – an average of $\mu = 537/576 \simeq 0.932$ hits per grid square. The number of hits in each of these grid squares was counted. The results are summarised in Figure 16.

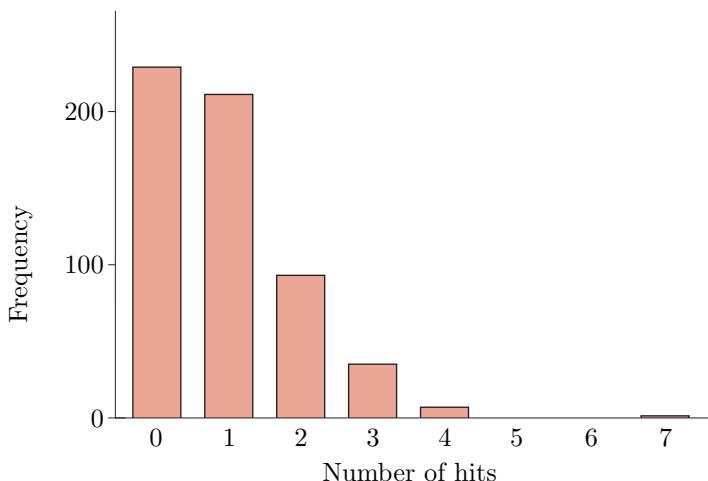


Figure 16 A bar chart for the number of hits in a grid square

As can be seen, many grid squares received no hits, almost as many had one hit, progressively fewer grid squares had two, three or four hits, and one grid square received seven hits.

The investigators looked at whether the distribution of the number of hits in a grid square was consistent with the hits being randomly located in the large 6 km by 6 km square.

A model was required for the number of hits in a grid square *assuming that the bombs landed in random positions*. As a first step towards developing such a model, each grid square was in turn subdivided into 900 ‘crater’ squares. As suggested by the name, each of these squares was roughly the size of the crater made by a V-1 bomb. No crater square within the 6 km by 6 km square of South London suffered more than one hit.

The following modelling assumptions were made.

- A hit at one point is no more likely than a hit at any other point. In other words, the probability that a crater square suffers a hit is constant from square to square.
- Where a bomb lands is not influenced in any way by where any other bomb lands. That is, whether or not a particular crater square is hit is independent of whether or not any other crater square is hit.
- There is a negligible chance of a bomb dropping into an existing crater. In other words, each crater square suffers at most one hit.

Together these assumptions mean that the outcomes (a hit or not a hit) for the 900 crater squares in a grid square may be regarded as 900 independent Bernoulli trials with constant probability of success (a hit). So since the average number of hits per grid square was 0.932, the number of hits in a grid square may be modelled by a binomial distribution with parameters $n = 900$ and $p = 0.932/900 \approx 0.00104$. For this model, the mean number of hits per grid square is $np = 0.932$, the mean number of hits observed. This is a situation involving a binomial distribution for which the parameter p is small and n is large.

In the activities in this chapter, you are asked to investigate the circumstances in which a binomial distribution may be approximated by a Poisson distribution. The first of these activities (Computer activity 33) concerns the modelling situation just described. You will need to use the animation **V-1 bombs** for all of these activities. If possible, do all three activities in one computer session.



Computer activity 33 Models for the V-1 data

- (a) As discussed in Example 1, if the positions where V-1 bombs landed were randomly distributed, then the number of hits in a grid square (when each grid square is divided into 900 ‘crater’ squares) can be modelled by the binomial distribution, $B(900, 0.932/900)$. In this part of the activity, you will use the animation **V-1 bombs** to compare the data with this model.

- Open the **V-1 bombs** animation.

The animation opens on a tabbed panel labelled **V-1 model**. This panel shows, on a single diagram, a bar chart of the V-1 data and the probability function of the binomial distribution $B(900, 0.932/900)$. The bar chart of the data is a version of the bar chart shown in Figure 16: here relative frequencies are used so that the heights of the bars sum to 1. Do you think that the bar chart of the data suggests that the V-1 bomb was an accurately aimed weapon?

- (b) When developing the model, each grid square was divided into 900 (30×30) crater squares. However, a different number of crater squares might have been chosen – either fewer or more than 900. If each grid square is divided into n crater squares (instead of 900), then the corresponding model for the number of hits in a grid square is $B(n, 0.932/n)$.

Underneath the plot of the bar chart and binomial probability model, there is a slider for the value of n . There is also a box giving the current value being used for n in the model shown in the plot: notice that the default value for n is 900 (because we have been considering $B(900, 0.932/900)$ initially). The slider for n is given on a log scale. This is to allow you to try both very small values of n (as small as 5), as well as very large values of n (as large as 2000).

- Move the slider for n (or type different values for n in its field and press **Enter**) to investigate how well the binomial model fits the data for different values of n .

For what range of values of n (approximately) does the binomial model appear to fit the data well?

Describe how the binomial probabilities change as n gets larger and larger.

- (c) In this final part, you will compare the binomial model for a large value of n with the Poisson(0.932) model.

- Set a very large value of n (2000, say). Observe what the probability function of the model looks like.
- Then click on the button labelled Poisson(0.932) to replace the binomial probability function on the diagram with the Poisson probability function.

Does the Poisson(0.932) model seem to fit the data well?

In Computer activity 33, you saw that for $\mu = 0.932$, a binomial distribution $B(n, \mu/n)$ may be approximated for large values of n by a distribution that depends only on the value of μ : this distribution is Poisson(μ). (Specifically, in Computer activity 33(c), you found that Poisson(0.932) was an excellent approximation to $B(2000, 0.932/2000)$.) What happens for other values of μ ? And how large must n be for the approximation to be a good one? You are asked to investigate these questions in Computer activity 34.

Computer activity 34 Good approximations



- Click on the **Binomial/Poisson** tab in the **V-1 bombs** animation.

The diagram shows two probability functions: one is for the binomial distribution $B(n, \mu/n)$, and the other is for the Poisson distribution with parameter μ , Poisson(μ). The default values of n and μ are the values from the model used for the V-1 bomb hits: 900 and 0.932, respectively. As can be seen from the animation, for these values of n and μ , the two probability functions are virtually indistinguishable on the diagram.

- Move the slider for n , first decreasing n , and then increasing n .

Notice that the diagram of the binomial probability function changes as n changes. You investigated these changes for $\mu = 0.932$ in Computer activity 33. For n greater than about 250, the two probability functions are almost indistinguishable. They differ more as n becomes smaller. But for n greater than about 50 they are still very close.

- Now change the value of μ to a value of your choice. Notice that both probability functions change automatically.
- By moving the slider for n , investigate how the binomial probabilities change with n .

For the value of μ that you chose, write down the (approximate) range of values of n for which you consider the Poisson distribution to be a good approximation to the binomial distribution.

- Repeat this investigation for a selection of different values of μ , some less than 0.932 and some greater than 0.932.

Try to summarise your conclusions briefly.

The binomial distribution is defined only for $\mu < n$, since the parameter p lies between 0 and 1, and $\mu = np$.

Notice that for $\mu \leq 0.7$, the vertical axis goes up to 1, compared with 0.5 for $\mu > 0.7$.

When using a binomial model, it is usual to specify the parameters n and p . Therefore, in order to be useful, any rule for the circumstances in which a Poisson approximation is good ought to be given in terms of n and p , rather than n and μ . In Computer activity 35, you are asked to investigate for what ranges of values of n and of p a Poisson distribution provides a reasonable approximation to a binomial distribution. Of course, what constitutes a ‘reasonable’ approximation is not clear-cut. An approximation may be sufficiently good for some purposes but not for others, depending on the accuracy required. By comparing diagrams of the probability functions, you are simply being asked to try to formulate a ‘rough rule’ for when a Poisson approximation appears to be good. For the example discussed in Section 1 of Unit 5, p was small and n was fairly large. Try to find out how small p needs to be and how large n should be for a Poisson approximation to be ‘good’.



Computer activity 35 More on approximations

- Click on the **Finding a rule** tab in the **V-1 bombs** animation.

The diagram shows two probability functions: one is for the binomial distribution $B(n, p)$, and the other is for the Poisson distribution $\text{Poisson}(np)$. The default values for n and p are 50 and 0.2, respectively. As you can see, for these values of n and p , although the shapes of the two distributions are similar, the Poisson distribution is not a very good approximation to the binomial distribution: the largest probabilities differ quite a lot.

- By using the slider for n below the plot, increase the value of n to around 200, while keeping the value of p set to 0.2. Is the Poisson distribution a good approximation to the binomial distribution after n has been increased to around 200?
- By varying the value of p (using the slider below the plot), investigate for what values of p the Poisson distribution is a good approximation to the binomial distribution when $n = 200$.
- Try to find a rough rule of your own involving n and p for when a Poisson distribution is a good approximation to a binomial distribution. You will shortly be able to compare this with ‘our’ rough rule (this is given in Section 1 of Unit 5).

In all these investigations you have probably compared only those binomial and Poisson probabilities that have the largest values. You may well not have compared probabilities in the tails of the distributions, where the probabilities are very small and discrepancies are difficult, if not impossible, to see. Even though, in the tails of the distributions, the absolute difference between a binomial and a Poisson probability may be extremely small, the percentage error in using the Poisson probability to approximate the binomial probability may be large. It follows that, even

when you believe a Poisson distribution to be a good approximation for a binomial distribution, you should be cautious about using a Poisson approximation to calculate probabilities in the tails.

10 Poisson processes

This chapter is associated with Subsection 3.2 of Unit 5.

In Subsection 10.1 you will be given several datasets to explore. Each dataset concerns events occurring in continuous time, and the purpose of your exploration will be to investigate whether or not a Poisson process might be a reasonable model for the occurrences of the events.

Subsection 10.2 consists of two activities in which you are asked to find probabilities associated with events occurring in a Poisson process, using Minitab to do the calculations.

10.1 Is a Poisson process a good model?

Data on the intervals between serious earthquakes worldwide have been discussed in Unit 5. In Computer activities 36 and 37, you are asked to explore these data using Minitab, and to reproduce some of the results and diagrams given in the unit.

Computer activity 36 Is the average rate constant?

The data on serious earthquakes are in the Minitab worksheet **serious-earthquakes.mwx**. Open this worksheet now.

The data are the intervals (in days) between successive serious earthquakes. In this activity you will investigate whether the rate at which serious earthquakes occurred remained constant over the period of observation. You will do this by plotting the number of earthquakes that have occurred so far against time.

- A serious earthquake occurred on 15 August 1950. The first entry in column **Interval** (186) is the number of days from then until the next serious earthquake. The second entry (77) is the number of days from that earthquake until the following one, and so on. Since the data are ordered, the waiting times may be added to give the times (in days after 15 August 1950) at which each serious earthquake occurred. So the first serious earthquake after the one on 15 August 1950 occurred after 186 days; the second serious earthquake occurred after 263 days ($186 + 77$); and so on. That is, one serious earthquake had occurred within 186 days of the one on 15 August 1950, two had occurred within 263 days, and so on.

Minitab provides a function called **Partial sum** that can be used to calculate these times. This is one of the functions available when using **Calculator...** from the **Calc** menu.

- Choose **Calc > Calculator** The **Calculator** dialogue box will open.
- To store the times in a variable named **Time**, type **Time** in the **Store result in variable** field. (The results will be stored in the first available column, which is **C2** in this case.)

Next you must enter the formula for calculating these times in the **Expression** field.

- Scroll through the list of functions until you come to **Partial sum**. Click on it to select it.
- Click on the **Select** button and **PARS(number)** will be entered in the **Expression** field. The word **number** will be highlighted.
- Type **C1** to replace **number** (to indicate that you want to find the partial sums of the numbers in column **C1**). Alternatively you could type **Interval**, or double-click on **Interval**. Note that Minitab puts quotes around the name of the variable when you double-click on the variable to input it into a function. You do not need to use quotes when typing in the name *unless* the name of the variable is two words such as the variable **Weight change** (which was used in Computer activity 14). In that case, you would need to type '**Weight change**'.
- Finally, click on **OK** and the partial sums will be stored in column **C2** of your worksheet.

- (b) Now create a column named **Number** giving the number of earthquakes that had occurred after each of the times in column **C2** – that is, a column containing the numbers $1, 2, \dots, 67$ in that order. The quickest way to do this is by using **Simple Set of Numbers....** (The use of **Simple Set of Numbers...** was described in Computer activity 30.)
- (c) Obtain a scatterplot of the number of earthquakes (**Y variables**) against time (**X variables**). You should obtain a diagram like the one in Figure 17.

Calc > Make Patterned Data > Simple Set of Numbers...

Graph > Scatterplot... and select **Simple**.

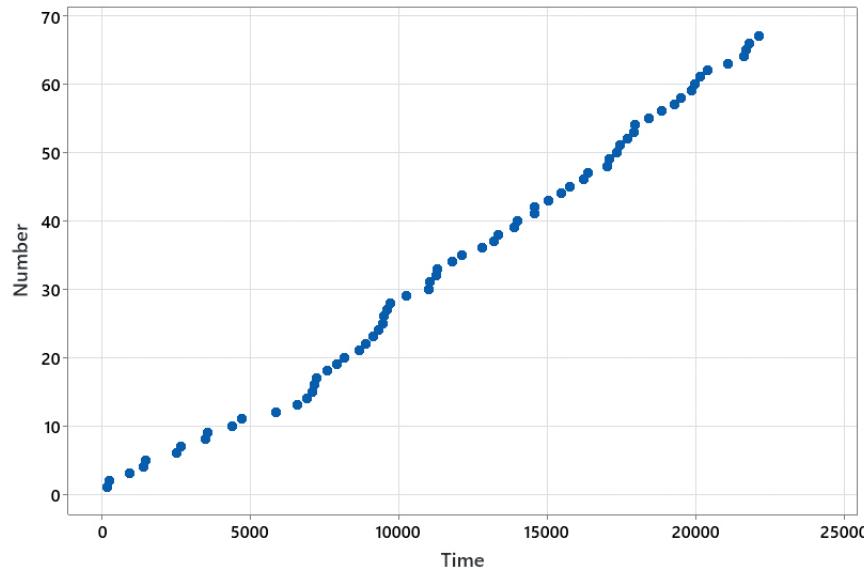


Figure 17 A scatterplot of number against cumulative time for the earthquakes data

As observed in the unit, the points lie quite close to a straight line through the origin, suggesting that the rate of occurrence of serious earthquakes remained constant over the period of observation.

Computer activity 37 Waiting times

The intervals, or waiting times, between successive events in a Poisson process are exponentially distributed. In this activity, you will explore the data in [serious-earthquakes.mwx](#) to see whether an exponential model is a good one for the intervals between serious earthquakes.

- Find the mean and standard deviation of the intervals between serious earthquakes. Check that these values are consistent with the data being observations from an exponential distribution.
- Obtain a histogram of the data with the following properties.
 - The ticks on the horizontal axis are at the cutpoints.
 - The bins have width 100.
 - The first bin starts at 0 and the last finishes at 1200.

Is the shape of the histogram consistent with an exponential distribution being a good model for the intervals between serious earthquakes?

Note that since waiting times cannot be negative, the first group must not extend below 0.

In Computer activity 36, you saw that the average rate of occurrence of serious earthquakes appears to have remained relatively constant over the period of observation; and in Computer activity 37, you found that an

exponential distribution seems to be a defensible model for the intervals between serious earthquakes. These results are consistent with a Poisson process being a reasonable model for the occurrences of serious earthquakes.

Computer activity 38 Coal-mining explosions

Jarrett, R.G. (1979) 'A note on the intervals between coal-mining disasters', *Biometrika*, 66(1), pp. 191–193.



On 24 October 1913 there was an explosion at the Universal Colliery, Senghenydd, South Wales, which was the worst disaster in British mining history. In total, 439 men were killed, one of whom was a distant relation of one of the authors of M248.

The Minitab worksheet **coal.mwx** contains historical data on the intervals in days between explosions in UK coal mines from 15 March 1851 to 22 March 1962, inclusive. There were 191 explosions altogether, including those on each of the two dates above. So the worksheet contains 190 waiting times. There is one zero: two explosions occurred on 6 December 1875.

In this activity, you should use the methods of Computer activities 36 and 37 to investigate whether a Poisson process might be a good model for the occurrences of coal-mining explosions.

- First, investigate whether an exponential distribution is a good model for the intervals between coal-mining explosions.
 - Find the mean and standard deviation of the waiting times between explosions.
 - Produce a histogram of the data with bins of width 100 and the first bin starting at 0 and the last bin ending at 2400.

Using your results, explain whether or not you think that an exponential model is a good one for the intervals between coal-mining explosions. (Follow the method outlined in Computer activity 37.)

- Produce a scatterplot of the number of explosions against time.
Do you think that the rate of occurrence of coal-mining explosions remained constant over the period of observation?
- Use your answers to parts (a) and (b) to explain whether or not you think that a Poisson process is a suitable model for the occurrences of coal-mining explosions during the period of observation.

You have now explored two datasets consisting of waiting times between events. You have seen that the data on times between serious earthquakes are reasonably consistent with a Poisson process being a good model for the occurrences of serious earthquakes. However, you found that there is good reason to doubt that the occurrences of coal-mining explosions may be modelled by a Poisson process: the rate of occurrence appears to have declined over the period of observation.

Of course, if the rate of occurrence of events remains constant over time, this does not necessarily mean that a Poisson process is a good model for the occurrences of the events. Consider, for instance, the simple situation where events occur at regular intervals: the events are completely predictable, and their rate of occurrence is constant. But if the rate

remains constant and the events are unpredictable, is a Poisson process necessarily a good model? In Computer activities 39 and 40 you are asked to investigate this using a dataset concerning the arrival of requests to perform a particular task. Try to work through both activities in the same Minitab session.

Computer activity 39 Requests to review

The results of academic research are reported in ‘papers’, which are published in collections called ‘journals’ after undergoing a process of ‘peer review’, that is, after review of their content by a small number of other experts in the field. Such reviews typically give rise either to publication of the paper after it has been revised, or to rejection of the paper as unsuitable for publication in the journal under consideration.

Suppose that the number of times that members of an academic department were asked to perform such reviews in each of 128 consecutive weeks were recorded to produce the data in the Minitab worksheet **review-requests.mwx**. (The first entry in the column **Requests** is for the first week, the second is for the second week, and so on.)

To produce a scatterplot of the cumulative number of requests (**Y variables**) against time (**X variables**), notice that the data in worksheet **review-requests.mwx** are slightly different to data in the worksheets so far considered in this chapter (**serious-earthquakes.mwx** and **coal.mwx**). In the latter two worksheets, the data were the *waiting times* between consecutive events, so **Partial sum** was used in Minitab to calculate the times (variable **Time**) at which the events occurred, and at each of these times the cumulative number of events that had occurred up to that time increased by one (variable **Number**). Here we have a different situation. In the worksheet **review-requests.mwx**, we are given the number of requests for each of $1, \dots, 128$ weeks. So we know that at time 1 (i.e. week 1), 4 events had occurred, at time 2 (i.e. week 2), $4 + 0 = 4$ events had occurred, and so on. Therefore create a scatterplot of number of requests against time as follows.

- Create a column named **Number** that contains the cumulative number of requests after each week by using **Calc > Calculator...** and entering **Number** in the **Store result in variable** field, and **PARS('Requests')** in the **Expression** field.
- Create a column named **Time** that contains the week numbers ($1, 2, \dots, 128$) by using **Calc > Make Patterned Data > Simple Set of Numbers...**
- Use the data in these two columns to produce the scatterplot.

What does this scatterplot tell you about the rate of occurrence of requests to review?



This method was used to find the variable **Time** in the previous activities.

This method was used to find the variable **Number** in the previous activities.

Graph > Scatterplot... and select **Simple**.

Computer activity 40 The number of requests in a week

You should still have the Minitab worksheet review-requests.mwx open.

If a Poisson process is a good model for the receipt of requests to review papers described in Computer activity 39, then the numbers of requests in a week are observations from a Poisson distribution. In this activity, you will investigate whether a Poisson distribution is a good model for the data.

- (a) Find the sample mean and sample standard deviation of the number of review requests in a week. What can you deduce from these values?
- (b) Since the numbers of review requests received each week are discrete, a bar chart should be used to get a feel for the distribution of the number of weekly requests. However, there is a problem in producing such a bar chart in Minitab.

The problem arises because the numbers of requests per week range from 0 up to 22, but not all of these numbers of requests were observed. For example, there were no weeks for which there were 14 or 15 requests, which means that there should be zero frequencies for these weeks. However, when drawing a bar chart, Minitab includes only categories that have non-zero frequency, which means that Minitab would produce a bar chart with no bars for 14 or 15 requests, and the bars for 13 and 16 would be adjacent. Hence a Minitab bar chart gives a misleading impression of the shape of the distribution of the data.

However, we can use a histogram to give an idea of the shape of the distribution of review requests. In order to make the histogram look like the corresponding bar chart, edit the scale so that the bins have width 1, while retaining **Midpoint** under **Interval Type** in the **Edit Bars** dialogue box (rather than changing it to **Cutpoint**). Obtain this histogram now, specifying the bins by entering **0:22/1** in the **Midpoint/Cutpoint positions** field of the **Binning** tab of the **Edit Bars** dialogue box.

- (c) A diagram of the probability mass function of a discrete probability distribution can be obtained using **Graph > Probability Distribution Plot...** Obtain a diagram of the probability mass function of a Poisson distribution with mean 4.016 (the value of the sample mean) as follows.
 - Choose **Graph > Probability Distribution Plot...** to obtain the **Probability Distribution Plots** dialogue box.
 - In the **Probability Distribution Plots** dialogue box, select **View Single** and click on **OK**.
 - In the **Probability Distribution Plot: View Single** dialogue box, choose **Poisson** from the **Distribution** drop-down list, and enter 4.016 in the **Mean** field.
 - Click on **OK** to obtain the diagram.

Compare your histogram of the data with the diagram of the Poisson probability mass function. Do you think that the Poisson distribution is a good model for the number of review requests in a week? Is a Poisson process a suitable model for the occurrences of review requests?

10.2 Probability calculations

Two distributions are used when calculating probabilities associated with events occurring in a Poisson process: the Poisson distribution and the exponential distribution. The number of events that occur in an interval of length t has a Poisson distribution with parameter λt , where λ is the rate of occurrence of the events per unit of time; and the waiting time between successive events has an exponential distribution with parameter λ .

As you have already seen for the binomial distribution, probabilities are found in Minitab using **Probability Distributions** from the **Calc** menu. When you select a family of distributions from the **Probability Distributions** submenu, a dialogue box will be opened. The only part of the dialogue box that varies from one family to another is the central part, where you specify which member of the family is required.

In Chapter 7, you saw that for a binomial distribution you must specify the **Number of trials**, which is the parameter n , and the **Event probability**, which is the parameter p .

As you will see when you try the activities in this section, for a Poisson distribution you must just specify the **Mean**; for the $\text{Poisson}(\mu)$ distribution, the mean is equal to the parameter μ .

You will also find that for an exponential distribution you must again specify the *mean* of the distribution. Take care when using **Exponential...**, however. Remember that the mean of an exponential distribution is *not* the same as the parameter: if the parameter is λ , then the mean is $1/\lambda$. Minitab actually requires you to enter values in two fields named **Scale** and **Threshold**. For an exponential distribution, the value in the **Threshold** field should be 0 (the default value), and the value in the **Scale** field should be the mean. Alongside the **Scale** field, Minitab reminds you that this quantity is the mean when the threshold equals zero. (Entering a non-zero value in the **Threshold** field specifies a distribution that has the same shape as an exponential distribution, but a different range. You will not need to use this.)

The following activities serve two purposes. First, they will give you the opportunity to ensure that you can use Minitab to find probabilities involving exponential distributions and Poisson distributions. Second, they will provide you with practice at deciding what probabilities are required in questions about Poisson processes.

Computer activity 41 Serious earthquakes

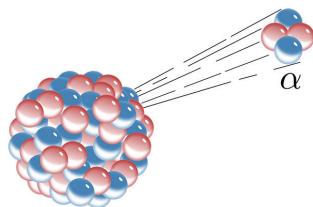
In Computer activities 36 and 37, you explored data on the waiting times, in days, between serious earthquakes covering the period from 15 August 1950 until 11 March 2011. The data appear to be consistent with a Poisson process being a reasonable model for the occurrences of serious earthquakes.

Suppose that the occurrences of serious earthquakes may be modelled by a Poisson process with rate $\lambda = 67/22\,120$ per day.

- Write down the distribution of the number of serious earthquakes that occur in a typical four-year period. Give the value of the parameter correct to three decimal places.
- Taking the approximate value for the mean in a four-year period that you found in part (a) as if it were exact, use Minitab to find each of the following probabilities.
 - The probability that exactly five serious earthquakes occur in a four-year period.
 - The probability that fewer than three serious earthquakes occur in a four-year period.
 - The probability that at least eight serious earthquakes occur in a four-year period.
- Write down the distribution of the waiting time in days between serious earthquakes. Calculate the mean waiting time to the nearest day.
- Taking the approximate value for the mean waiting time that you found in part (c) as if it were exact, use Minitab to find the probability that the gap between successive serious earthquakes
 - will be less than 30 days
 - will exceed two years.

Computer activity 42 Emissions of alpha particles

In Example 13 of Unit 5, you saw that the Poisson distribution is a good fit for the data given on the numbers of particles emitted from a radioactive source in $7\frac{1}{2}$ -second intervals.



Suppose that emissions of alpha particles may be modelled by a Poisson process with rate $\lambda = 0.517$ per second. (This is the estimate for λ obtained from the data.)

- Write down the distribution of the number of alpha particles emitted in a one-minute period.
- (i) Find the probability that exactly 30 particles are emitted in a one-minute period.

- (ii) Find the probability that fewer than 20 particles are emitted in a one-minute period.
- (iii) Find the probability that at least 50 particles are emitted in one minute.
- (c) Write down the probability distribution of the interval between successive emissions of particles.
- (d) (i) Find the probability that the interval between successive emissions is less than half a second.
(ii) Find the probability that the interval between successive emissions exceeds three seconds.

11 Quantiles of continuous distributions



This chapter is associated with Subsection 4.1 of Unit 5.

In Subsection 4.1 of Unit 5, you were introduced to the idea of quantiles of continuous distributions. In this chapter, you will consider these further using the animation **Quantiles**.

Computer activity 43 The link between α and q_α



- Open the **Quantiles** animation.

The animation opens on a tabbed panel labelled **Symmetric distribution**. This panel shows two linked graphs.

The top graph is the c.d.f. of a continuous distribution, and looks rather like Figure 11(b) from Unit 5. The α -quantile, q_α , is marked on the horizontal axis of the graph, while the corresponding value of α is marked on the vertical axis. By definition,

$$F(q_\alpha) = \alpha,$$

and so, for any given value of α ($0 < \alpha < 1$), the c.d.f. can be used to find the associated value q_α . This is illustrated pictorially on the top graph by the dotted line linking the value of α with q_α through F .

The bottom graph is the p.d.f. of the distribution, and looks rather like Figure 12 from Unit 5. As for the top graph, q_α is marked on the horizontal axis: this is the *same* q_α as on the top graph. Now, because

$$F(q_\alpha) = P(X \leq q_\alpha),$$

it follows that

$$P(X \leq q_\alpha) = \alpha.$$

So on the bottom graph, the value of α is the area under the p.d.f. to the left of q_α . This is illustrated pictorially on the bottom graph by the shaded

This is actually the c.d.f. of a distribution known as the standard normal distribution, which will be discussed in detail in Unit 6.

area. Therefore, if a value is given for α ($0 < \alpha < 1$), then this value specifies the size of the shaded area, and hence the value of q_α .

Underneath the two graphs is a slider bar to change the value of α . Which quantile is being shown on the graphs initially?

- Use the slider for α to decrease the value of α down towards 0, and then increase the value of α up to 1. Notice that the value of α will change on both graphs simultaneously.

Comment on what you observe.

The next activity considers quantiles for two different continuous distributions.



Computer activity 44 Quantiles of two other continuous distributions

- Click on the **Exponential distribution** tab of the animation **Quantiles**.

As for the **Symmetric distribution** panel, this panel has two linked graphs. The top graph is the c.d.f. of the exponential distribution with parameter λ equal to 1, and the bottom graph is the p.d.f. for this distribution. Once again, each graph has q_α marked on the horizontal axis. The top graph illustrates pictorially how the value of α is linked to q_α through the c.d.f. F in the exponential distribution case, and the bottom graph illustrates pictorially how the value of α , which determines the size of the shaded area under the p.d.f., is linked to the value of q_α .

Although the c.d.f. and p.d.f. of the exponential distribution look quite different to the c.d.f. and p.d.f. of the distribution on the previous panel, the idea of quantiles, and how they are calculated for a given value of α , works in exactly the same way for both distributions.

- Decrease and then increase the value of α by moving the slider for α (found underneath the graphs).

When changing the value of α you should notice that the value of q_α changes in the same way as it did for the previous distribution: that is, as α decreases, so does q_α , and as α increases, so does q_α .

- Click on the **Power distribution** tab of the animation **Quantiles**.

Again, this panel shows two linked graphs. This time, the graphs show the c.d.f. and p.d.f. of a power distribution, with p.d.f. given by

$$f(x) = 3x^2, \quad 0 < x < 1,$$

and, for $0 < x < 1$, c.d.f. given by

$$F(x) = x^3.$$

For this distribution, both the c.d.f. and p.d.f. look very similar, although they are, of course, different functions.

As for the other two panels, the two graphs illustrate pictorially how the value of α is linked to q_α through the c.d.f. (top graph) and p.d.f. (bottom graph).

- Decrease and increase the value of α by moving the slider for α (found underneath the graphs).

You should notice, once again, that as α decreases, so does q_α , and as α increases, so does q_α .

12 Calculating quantiles

This chapter is associated with Subsection 4.2 of Unit 5.

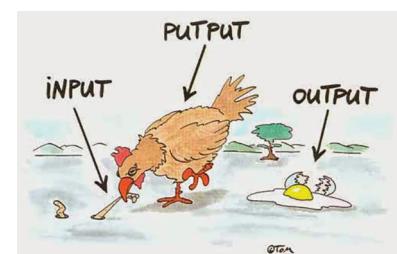
In Minitab, you have used **Probability Distributions** from the **Calc** menu to find probabilities for binomial, Poisson and exponential distributions. In this chapter, you will use **Probability Distributions** to find quantiles for a number of distributions, both continuous and discrete.

When you choose a family of distributions from the **Probability Distributions** submenu, a dialogue box opens. Whichever family you choose, this box requires similar information. Quantiles are values of the inverse of the cumulative distribution function; so to find quantiles, you must select **Inverse cumulative probability** from the top section of the dialogue box. In the middle section of the dialogue box, you must specify which member of the family of distributions is required. The bottom section of the box is where you specify the input and (optionally) where to store the results.

When finding quantiles, the values that you input (using either **Input column** or **Input constant**) must be decimal numbers between 0 and 1: they are the values of α for which the quantiles q_α are required.

The first activity illustrates the use of Minitab to find quantiles for a continuous random variable.

So far we have used only **Probability (or Probability density)** and **Cumulative probability**.



Computer activity 45 Quantiles of an exponential distribution

In Computer activity 41, an exponential distribution with mean 330 was used to model the waiting time in days between successive serious earthquakes worldwide. In this activity you will use Minitab to find the median, quartiles and deciles of this distribution.

(a) Follow the instructions below to find the median.

- Obtain the **Exponential Distribution** dialogue box.
- Select **Inverse cumulative probability**.
- Enter 330 in the **Scale** field (and 0 in the **Threshold** field).

Calc > Probability Distributions > Exponential...

Recall that the median m is $q_{0.5}$.

- To find the median waiting time between serious earthquakes, select **Input constant** and enter 0.5 in the **Input constant** field.
- Click on **OK** and you will obtain the following output in the Session window.

Exponential with mean = 330

$P(X \leq x)$	x
0.5	228.739

This says that $P(X \leq 228.739) = 0.5$, that is, $F(228.739) = 0.5$, so the median waiting time between serious earthquakes is approximately 229 days (as you found in Activity 20 of Unit 5).

Now find the lower quartile q_L and the upper quartile q_U of the waiting time between serious earthquakes, and hence find the interquartile range of the waiting time between serious earthquakes.

- The simplest way to find all the deciles (that is, $q_{0.1}, q_{0.2}, \dots, q_{0.9}$) is to enter the values 0.1, 0.2, ..., 0.9 in a column of the worksheet, and then type the column name in the **Input column** field in the **Exponential Distribution** dialogue box. First open a new blank worksheet, then enter the numbers 0.1, 0.2, ..., 0.9 in column C1. Use **Exponential...** to store the deciles $q_{0.1}, q_{0.2}, \dots, q_{0.9}$ in column C2.
- In part (b), you found that $q_{0.2}$ is approximately 74 and $q_{0.6}$ is approximately 302. So, according to the model, approximately 20% of intervals between serious earthquakes are shorter than 74 days, and approximately 40% of intervals exceed 302 days. Interpret the deciles $q_{0.1}$ and $q_{0.8}$.

The next two activities illustrate the use of Minitab to find quantiles for discrete random variables. Recall from Subsection 4.2 of Unit 5 that quantiles of a discrete random variable X are values in the range of X . So, for instance, if X takes integer values, then the quantiles must be integers.

Computer activity 46 Quantiles of a binomial distribution

In Computer activities 27 to 29, you used Minitab to find various probabilities associated with a multiple choice examination consisting of twenty questions. Each question had five options, exactly one of which was correct. If X is a random variable representing the number of questions answered correctly by a student who guesses answers at random, then X has a binomial distribution: $X \sim B(20, 0.2)$. In this activity, you will find several quantiles for this distribution.

- (a) Use a procedure similar to that used for the median of an exponential distribution in Computer activity 45 to find the median score of students who guess answers at random. You should obtain the following output in the Session window.

Use **Calc > Probability Distributions > Binomial...** and enter 20 and 0.2 for the parameters of the distribution.

Binomial with $n = 20$ and $p = 0.2$

x	$P(X \leq x)$	x	$P(X \leq x)$
3	0.411449	4	0.629648

Values of the cumulative distribution function are given for two consecutive values of x in the range of X . For the first of these, the value of the c.d.f. is less than 0.5: $P(X \leq 3) = 0.411449$. For the second, the c.d.f. takes a value greater than 0.5: $P(X \leq 4) = 0.629648$. The median m is defined to be the smallest value of x in the range of X for which $F(x) \geq 0.5$. So the median score of students who guess answers at random is the second of the two displayed values of x ; that is, the median score is 4.

Find the lower quartile and the upper quartile of X .

- (b) The pass mark is to be set so that only one in a thousand students who guess answers at random will pass. What quantile of X should you find in order to determine what the pass mark should be? What should the pass mark be?

Notice that when finding quantiles for a binomial distribution in Computer activity 46, Minitab displayed two probabilities: the first probability gave the value of x for which $P(X \leq x) < \alpha$, and the second gave the value of x for which $P(X \leq x) > \alpha$. Using the definition of a quantile for discrete distributions given in Unit 5, the quantile is then the second of these displayed values. Minitab often gives two values like these when using **Inverse cumulative probability** for discrete probability distributions, and the same principle applies for identifying the quantiles for other discrete distributions. In the final activity in this chapter, you are asked to find several quantiles for Poisson distributions.

Computer activity 47 Quantiles of Poisson distributions

- (a) In Computer activity 41, a Poisson distribution with parameter 4.425 was used to model the number of serious earthquakes that occur in a typical four-year period. According to this model, what is the median number of serious earthquakes that occur in a four-year period?

- (b) Telephone calls arrive at a switchboard at random at an average rate of 40 per hour. The number of calls that arrive in an hour may be modelled by a Poisson distribution with parameter 40.
- What is the largest number of calls received during any of the 10% of hours that are the quietest?
 - What is the largest number of calls received during any of the 1% of hours that are the quietest?

Computer exercises

Computer exercise 1 Bronze Age cups

Baxter, M.J. (1984) *Exploratory Multivariate Analysis in Archaeology*. Edinburgh University Press.



Pottery from the Apennine culture has been found on the Capitoline Hill, one of the seven hills of Rome

This exercise concerns some archaeological data, on the dimensions of Middle Bronze Age cups from the Apennine culture of what is now central and southern Italy. The data are in the Minitab worksheet **cups.mwx**. The two columns of the worksheet contain linked data on the rim diameters and heights (in cm) of $n = 60$ such cups.

- Obtain a unit-area histogram of the rim diameters of the cups, using cutpoints starting at 6 cm, ending at 30 cm, in bins of width 2 cm. What does this histogram tell us about the distribution of rim diameters?
- Obtain a scatterplot of height (on the vertical axis) against rim diameter (on the horizontal axis) for these cups. What does the scatterplot tell you about the relationship between height and rim diameter?
- Given the solutions to parts (a) and (b), can you make a conjecture as to what features the distribution of heights might have? Check out your guess by obtaining a unit-area histogram of the heights of the cups, using cutpoints starting at 3 cm, ending at 15 cm, in bins of width 1 cm.

Computer exercise 2 Comparing industries

Data downloaded in March 2016 from Bureau van Dijk.

The Minitab worksheet **industries.mwx** contains information about 55 manufacturing companies in Europe in 2014. Each of these companies has been classified into one of three industry groups indicating what the company's economic activity is focused on. The groups are coded '16', '18' and '19' and, for simplicity, we will refer to the groups by these codes throughout. They actually stand for:

- 16 Manufacture of wood and of products of wood and cork (except furniture), and manufacture of articles of straw and plaiting materials
- 18 Printing and reproduction of recorded media
- 19 Manufacture of coke and refined petroleum products.

The group code is given in the variable `Industry code`. In addition, the number of employees in each company is given in the variable `Employees`.

- Obtain a (horizontal) comparative boxplot of the numbers of employees for manufacturers in each of the three industry groups.
- Describe how manufacturers in these three industry groups differ in the numbers of employees that they have.

Computer exercise 3 Passengers on standby

According to the publicity department of an international airline, on average 10% of people making reservations on their scheduled London–Rome service fail to turn up for their flight. Suppose that on a particular day, the flight is fully booked for 140 passengers with 16 more waiting on standby.

- If the airline's claim is accurate, what is the distribution of X , the number of passengers who do not turn up for the flight?
- Find the probability that all the standby passengers get seats on the flight.
- Find the probability that exactly half of the standby passengers get seats on the flight.



Computer exercise 4 Admissions to an intensive-care unit

The arrival times of patients at an intensive-care unit were recorded in the period 4 February 1963 to 19 March 1964. The Minitab worksheet `intensive-care.mwx` contains data on the 40 waiting times (to the nearest half hour) between the first 41 admissions. The purpose of collecting the data was to identify any systematic variations in arrival rates.

- Investigate whether or not these data are consistent with an exponential distribution being a good model for the waiting times between admissions.
- The data are ordered. Investigate whether the data are consistent with the admission rate remaining constant over time.

Cox, D.R. and Snell, E.J. (1981) *Applied Statistics*. London: Chapman and Hall. (The data were collected by Dr A. Barr, Oxford Regional Hospital Board.)

Computer exercise 5 Major explosive volcanic eruptions

This is the model that was proposed in Exercise 5 of Unit 5.



The people of Catania in Sicily are extremely interested in the occurrences of eruptions of Mount Etna, one of the most active volcanoes in the world, that towers above them

Suppose that the occurrences of major explosive volcanic eruptions in the northern hemisphere may be adequately modelled by a Poisson process with rate $\lambda = 0.0352$ per month.

- (a) Find the probability that there will be more than five such eruptions in a ten-year period.
 - (b) Find the probability that the waiting time between successive eruptions will be less than six months.
 - (c) Find the value x such that only 5% of intervals between successive eruptions are shorter than x months.
 - (d) Find the value y such that only 1% of intervals between successive eruptions exceed y years.
-

Solutions to computer activities

Solution to Computer activity 9

Obtain the **Edit Bars** dialogue box, as described in Computer activity 8. In the **Binning** panel, select **Cutpoint** and enter $15:29/2$ in the **Midpoint/Cutpoint positions** field. Figure S1 should be produced.

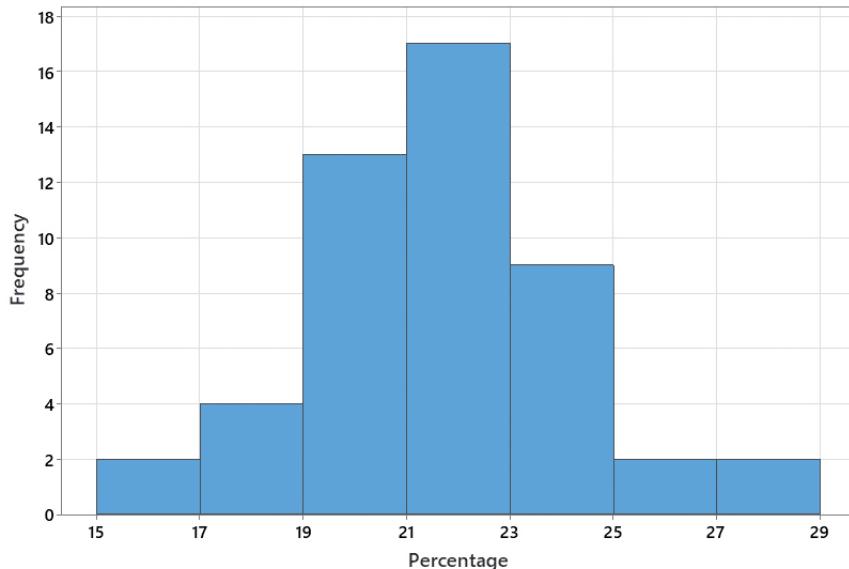


Figure S1 A further histogram of the sports club membership data

Solution to Computer activity 10

- (a) The histogram shown in Figure S2 is obtained by producing a histogram of the variable **Wage** and then selecting **Cutpoint** in the **Binning** panel of the **Edit Bars** dialogue box and entering $100:160/5$ in the **Midpoint/Cutpoint positions** field.

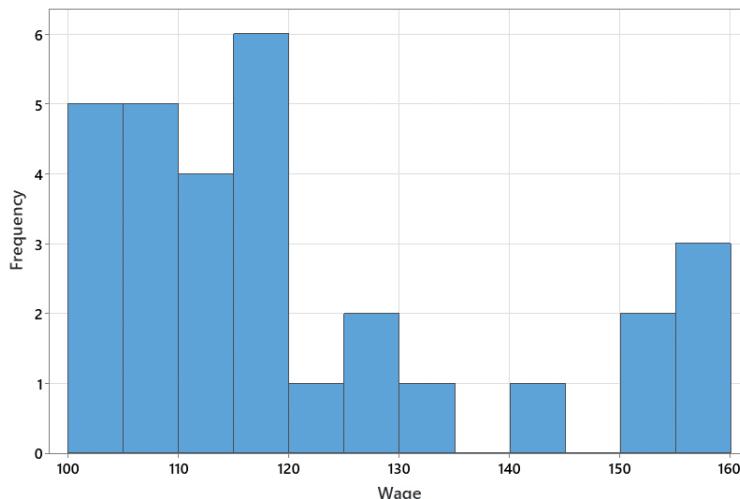


Figure S2 A histogram of the US wages data

- (b) The histogram is multimodal. In particular, it suggests that the production-line workers are split into at least two identifiable groups on the basis of their wages. The main group is towards the lower end of the wage range, while there appears to be a second, smaller group at the top end of the wage range.

Solution to Computer activity 11

The main difference between the boxplot in Figure 8 and the boxplot in Figure 13 of Unit 1 is that Minitab has drawn the boxplot vertically rather than horizontally. Apart from that, the positions of the ‘box’ and the ‘whiskers’ are shown in the same way as in Unit 1.

Solution to Computer activity 14

- (a) To obtain the required numerical summaries, in the **Display Descriptive Statistics: Statistics** dialogue box, you should have selected the following options: **Mean, Standard deviation, Median, Interquartile range, N nonmissing, N missing**.

The numerical summaries given by Minitab are:

Statistics

Variable	N	N*	Mean	StDev	Median	IQR
Weight change	81	2	-0.244	1.524	-0.200	1.650

Notice that because the value for **N*** is not equal to zero, this means that for some of the participants (two of them, in fact) weight change during the first two weeks is not available for some reason.

To edit the vertical boxplot so that the boxplot is displayed horizontally, you should have followed the procedure in the second half of Computer activity 12, as follows.

- Double-click on the boxplot to bring up the **Edit Graph** panel.
- Select the vertical axis on the boxplot, then double-click on it (or press **Ctrl+T**) to open the **Edit Scale** dialogue box.
- In the **Scale** panel of the **Edit Scale** dialogue box, select **Transpose value and category scales**.
- Click on **OK**.

The boxplot you should have obtained is shown in Figure S3.

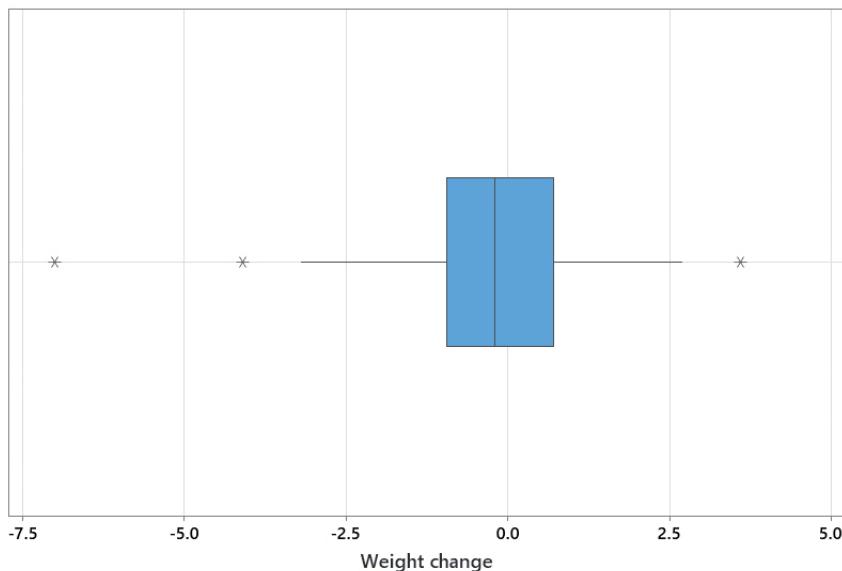


Figure S3 Boxplot of weight change for the response inhibition training clinical trial

- (b) One thing that stands out on the boxplot is that there are three potential outliers: two participants had particularly large weight losses, and one participant had a particularly large weight gain. Apart from these individuals, the data appear to be reasonably symmetric. The two whiskers are roughly the same length, and the two halves of the box also appear to be similar.
- (c) As the data contain potential outliers, the median and interquartile range are good choices to summarise these data. The median weight change is -0.200 kg and the interquartile range is 1.650 kg . This means that the typical participant did lose a little weight in the first two weeks of the clinical trial.

Solution to Computer activity 15

The numerical summaries provided by Minitab are:

Statistics

Variable	Group	N	N*	Mean	StDev	Median	IQR
Weight change	0	41	1	0.171	1.201	0.100	1.700
	1	40	1	-0.670	1.708	-0.550	1.700

Yes, there does appear to be a difference between the two groups. The participants in the treatment group (Group 1) on average lost weight (0.55 kg using the median and 0.67 kg using the mean) whereas the participants in the control group (Group 0) gained weight slightly (0.10 kg

using the median and 0.17 kg using the mean). The spread of values in the two groups is not notably different; in particular, the interquartile ranges are identical.

Solution to Computer activity 16

The required bar chart is shown in Figure S4.

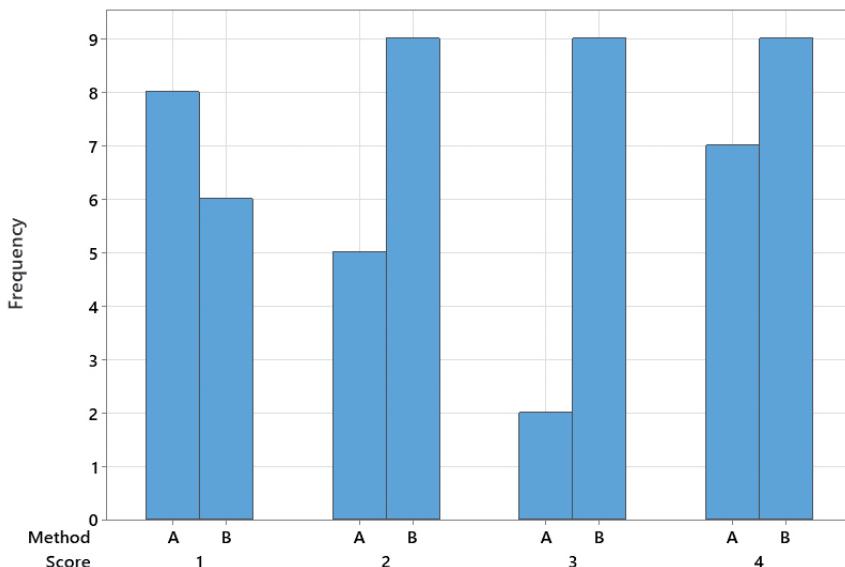


Figure S4 Quality of removal for different methods of tattoo removal

This was obtained by following the procedure given in the activity, except for entering **Score Method** (instead of **Score Depth**) in the **Categorical variables (2-4, outermost first)** field of the **Bar Chart: Counts of unique values, Cluster** dialogue box. From Figure S4, it seems that Method B gives each quality of removal score equally often except for the poorest score, which occurs less often, while Method A usually either works very well or poorly, but much less often gives intermediate results.

Solution to Computer activity 18

- First, a unit-area histogram of weight change for any set of cutpoints needs to be created. This is done by entering **Weight change** in the **Graph variables** field in the **Histogram: Simple** dialogue box. Then, on the **Y-Scale Type** tab of the **Histogram: Scale** dialogue box, make sure that the **Density** option is selected.

Graph > Histogram... and select **Simple**.

On the resulting histogram, alter the cutpoints used by doing the following. On the **Binning** tab of the **Edit Bars** dialogue box (obtained by selecting and double-clicking on the bars in the histogram), make sure that **Cutpoint** is selected and enter $-8:4/1$ in the **Midpoint/Cutpoint positions** field. The final histogram is shown in Figure S5.

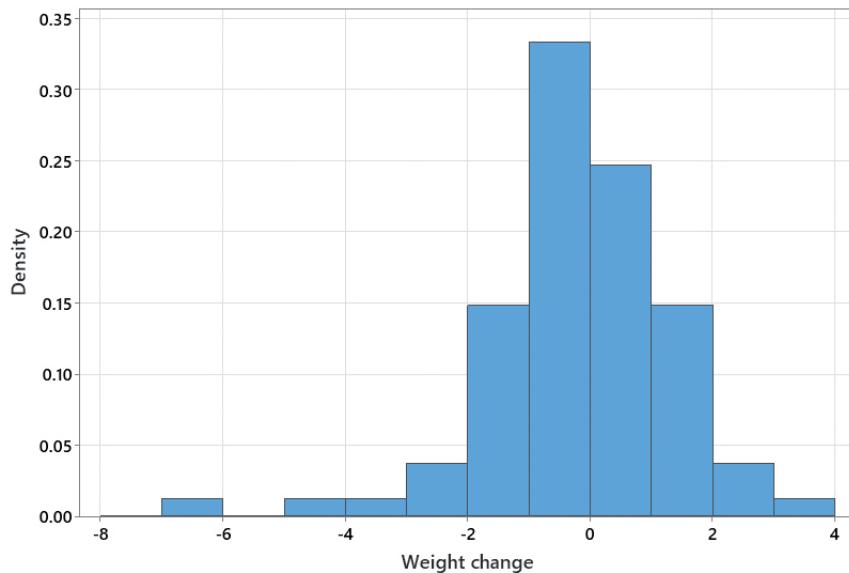


Figure S5 Unit-area histogram of weight change data

- (b) There is one clear peak on the histogram, so the data appear to be unimodal. There is a longer tail to the left-hand side of the histogram than to the right, suggesting that the data are left-skew.

Solution to Computer activity 19

- (a) In the worksheet, all the values for the weight changes are given in one column (**Weight change**) and there is a different column (**Group**) which indicates whether the participant was in the treatment or control group. (The treatment group is labelled with 1, the control group with 0.)
- (b) The comparative boxplot is shown in Figure S6.

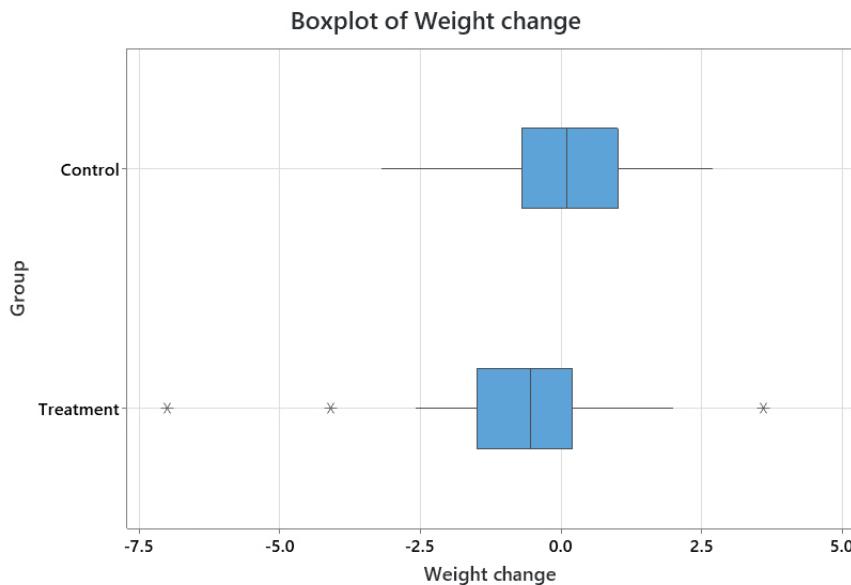


Figure S6 Comparative boxplot of weight change data, by treatment group

- (c) The boxplots indicate that the loss of weight was greater ('more negative') in the treatment group compared with the control group, as generally the boxplot for the treatment group is to the left of the boxplot for the control group. In particular, more than 50% of the treatment group lost weight whereas more than 50% of the control group gained weight. However, the change in weight appears to have been less consistent in the treatment group. In that group there are three potential outliers (including the participant who gained the most weight), whereas there are no outliers in the control group.

Solution to Computer activity 20

The comparative boxplot is shown in Figure S7.

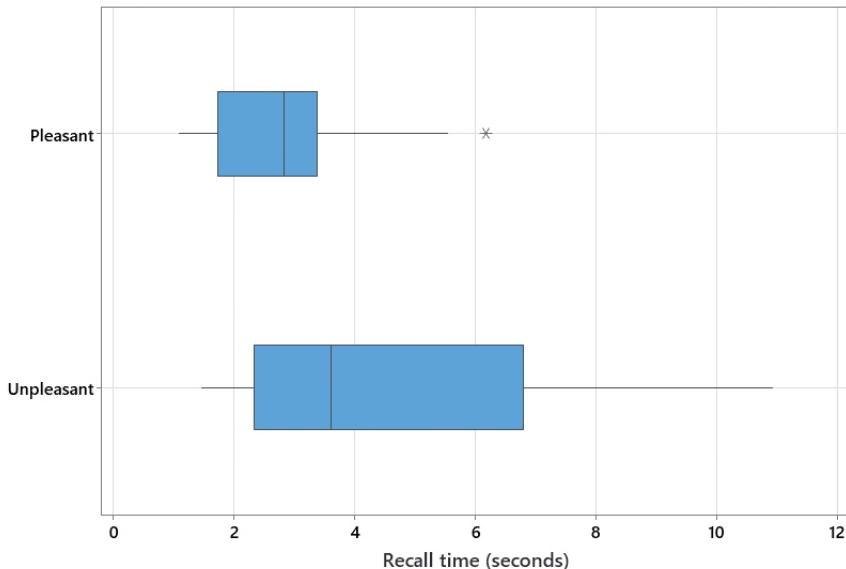


Figure S7 Comparative boxplot of memory recall data, by type of memory

Solution to Computer activity 21

- (a) In the **Boxplot: One Y, With Groups** dialogue box, enter **Weight change** in the **Graph variables** field and **Group Gender** in the **Categorical variables for grouping (1-4, outermost first)** field. On the **Axes and Ticks** tab of the **Boxplot: Scale** dialogue box, make sure that the **Transpose value and category scales** option is selected.

On the resulting boxplot, the labels for the categorical variables can be changed by editing the **Tick Labels** section of the **Labels** tab on the **Edit Scale** dialogue box (which is obtained by double-clicking on the vertical axis). Notice that **Auto** needs to be deselected before you can change the labels.

Graph > Boxplot... and select **With Groups under **One Y**.**

The edited comparative boxplot is shown in Figure S8.

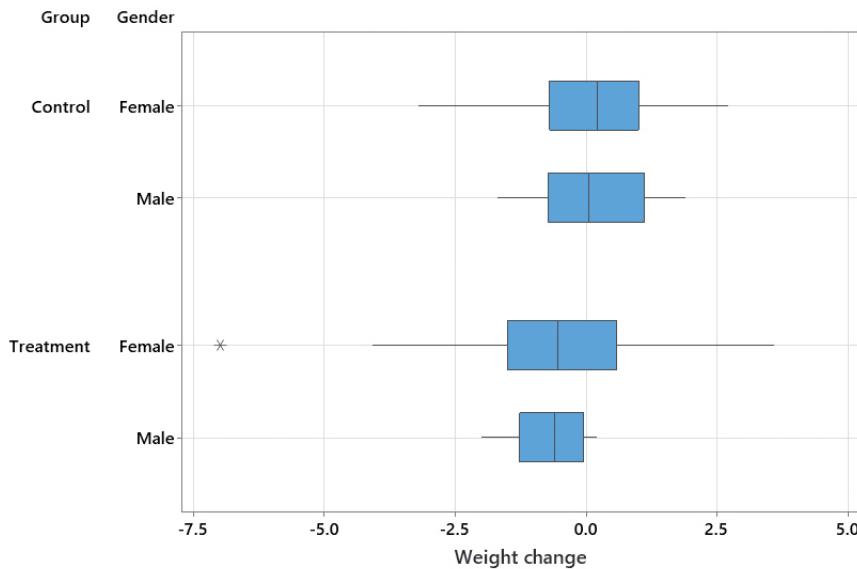


Figure S8 Comparative boxplot of weight change data, by treatment group and gender

- (b) There does not appear to be a difference between the ‘average’ weight changes experienced by females and males in the clinical trial: within each group, the locations of the boxes appear to be similar. The weight changes for females are, however, more spread out than those for males, particularly in the treatment group.

Solution to Computer activity 23

- (a) The basic scatterplot is produced by entering Road under **Y variables** and Map under **X variables** in the **Scatterplot: Simple** dialogue box. After the labels have been altered, and the title deleted, the scatterplot is as shown in Figure S9.

Graph > Scatterplot... and select **Simple**.

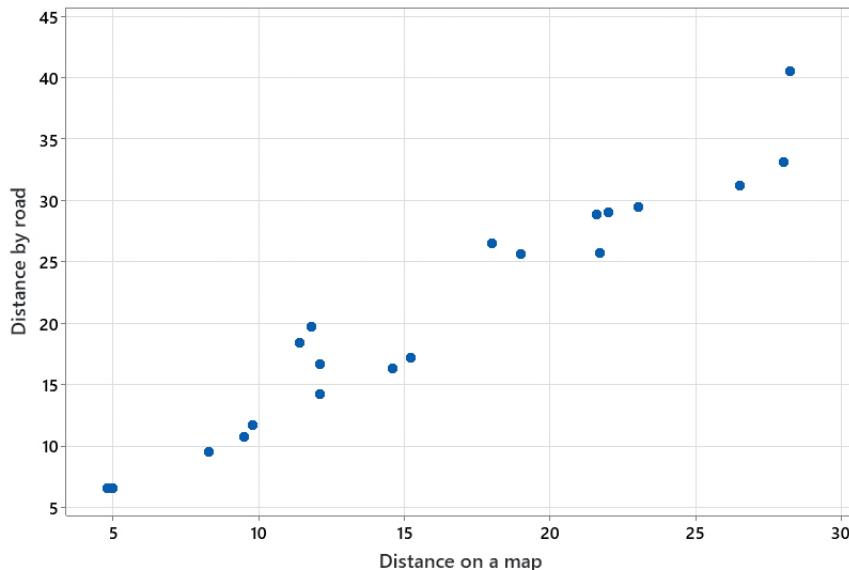


Figure S9 Road distances and map distances

- (b) There appears to be quite a strong positive linear relationship between the map distances and the distances by road. This is because the points appear to lie quite close to a straight line sloping upwards. None of the points appear to be much out of line with others, suggesting that there are no outliers. (The slope at which any line fitted to these data increases would give an estimate of the factor by which road distances are greater than map distances. It turns out that, based on these data, road distances are typically about 25% or so longer than map distances.)

Solution to Computer activity 29

- (a) The probability required is

$$P(T < 4) = P(T \leq 3) = F(3).$$

Calc > Probability
Distributions > Binomial...

Using **Cumulative probability** and setting **Input constant** equal to 3 gives $0.411449 \simeq 0.411$.

- (b) The probability required is

$$\begin{aligned} P(4 \leq T < 10) &= P(4 \leq T \leq 9) \\ &= P(T \leq 9) - P(T < 4) = F(9) - F(3) \\ &= 0.997405 - 0.411449 = 0.585956 \simeq 0.586. \end{aligned}$$

Notice that Minitab does not provide a facility to evaluate this probability directly.

Calc > Probability
Distributions > Binomial...

Solution to Computer activity 31

- (a) If X is the number of questions out of 10 that are answered correctly, then $X \sim B(10, 0.125)$. So in the **Binomial Distribution** dialogue box, you need to set **Number of trials** equal to 10 and **Event probability** equal to 0.125.

The probability that a student just passes is $P(X = 5)$. Using **Probability** and setting **Input constant** equal to 5 gives

$$P(X = 5) = 0.0039445.$$

So the probability that a student who guesses answers at random just passes the test is approximately 0.0039.

The probability that a student fails is $P(X \leq 4)$. Using **Cumulative probability** and setting **Input constant** equal to 4 gives

$$P(X \leq 4) = 0.995545.$$

So the probability that a student who guesses answers at random fails the test is approximately 0.9955.

- (b) If X is the number of questions out of 30 that are answered correctly in this test, then $X \sim B(30, 0.25)$. So you need to set **Number of trials** equal to 30 and **Event probability** equal to 0.25.

The probability that a student who guesses answers at random passes the test is

$$P(X \geq 15) = 1 - P(X \leq 14).$$

Using **Cumulative probability** and setting **Input constant** equal to 14 gives

$$P(X \leq 14) = 0.997250.$$

So the probability that a student who guesses answers at random passes the test is

$$P(X \geq 15) = 1 - 0.997250 = 0.002750 \simeq 0.0028.$$

Solution to Computer activity 32

- (a) You should have found that the bar chart for the sample of size 50 is the most jagged of the three, then that for the sample of size 500. The heights of the bars on the chart for the sample of size 5000 are very similar. The relative frequencies of the months seem to be settling down as the sample size increases.

- (b) For samples of size 82, you should have found that several, but not necessarily all, of the bar charts obtained were pretty much as jagged as the bar chart of the data. Figure S10 shows screenshots of three bar charts generated by the animation for samples of size 82.

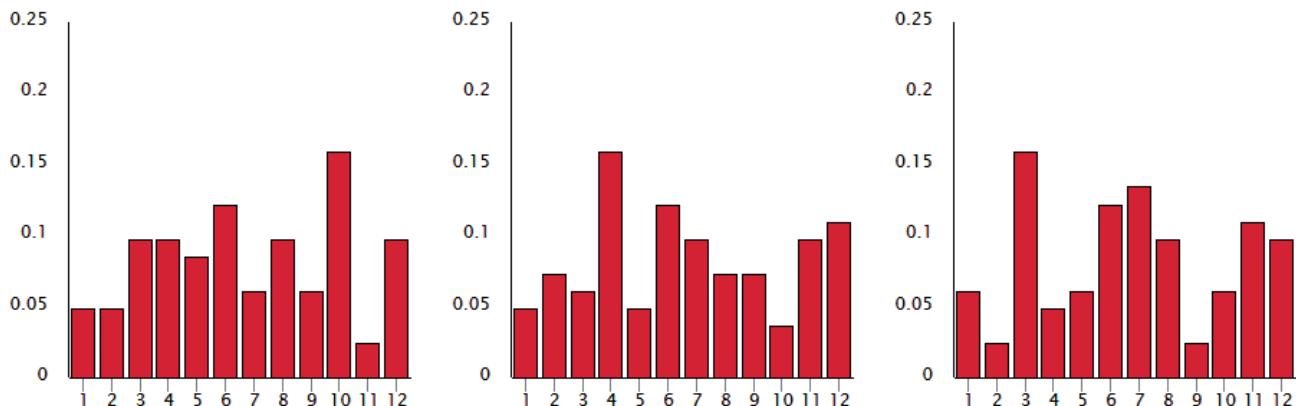


Figure S10 Bar charts of discrete uniform simulated samples of size 82

It is therefore possible that the variation in the numbers of deaths occurring in different months is just due to chance, and it is plausible that deaths are as likely to occur in any particular month as in any other. It can be argued that a discrete uniform distribution appears to be a reasonable model here.

Solution to Computer activity 33

- (a) The model appears to fit the data remarkably well. So there is no reason to suppose that the bombs did not land randomly in the 6 km by 6 km square: it does not look as though the V-1 bomb was an accurately aimed weapon.
- (b) The binomial model $B(n, 0.932/n)$ appears to fit the data well for values of n greater than about 50. The binomial model fits the data less and less well as n becomes smaller; and the fit is not good for n less than about 20.

For n greater than 50, the binomial probabilities change very little as n changes.

- (c) The change from binomial probability function to Poisson probability function is not noticeable for large enough n . (Yes, the animation is working: observe a change when you compare the Poisson model with a binomial model using, say, $n = 50$ or smaller.) The Poisson model also fits the data well.

Solution to Computer activity 34

For relatively large values of μ (for $\mu > 5$, say), the Poisson distribution is a good approximation to the binomial distribution only for very large values of n . For smaller values of μ , the Poisson distribution is a good approximation for a wider range of values of n .

Solution to Computer activity 35

- (a) When $p = 0.2$, the Poisson distribution still doesn't seem to be a very good approximation for the binomial distribution when n has increased to 200.
- (b) As the value of p is reduced, the Poisson distribution provides a better and better approximation to the binomial distribution. When p is around 0.07 or smaller, the approximation is quite good for $n = 200$.
- (c) You will shortly be able to compare your rough rule with one given in Unit 5. However, whatever your rule, the bombing situation described in Example 1 is likely to satisfy it: p is very small ($p \simeq 0.001$) and n is large ($n = 900$).

Solution to Computer activity 36

- (b) A column containing the numbers $1, 2, \dots, 67$ can be created as follows.
 - Choose **Calc > Make Patterned Data > Simple Set of Numbers...** to obtain the **Simple Set of Numbers** dialogue box.
 - To store the numbers in a column called **Number**, type **Number** in the **Store patterned data in** field.
 - All the integers from 1 to 67 are required, so enter 1 in the **From first value** field and 67 in the **To last value** field.

The other fields should each contain default values of 1. (If by any chance they do not, then change the values in these fields to 1.)

- Click on **OK**.

The numbers 1 to 67 will be stored in column C3 (the first available column in the worksheet).

Solution to Computer activity 37

- (a) The mean and standard deviation can be found as follows.
 - Obtain the **Display Descriptive Statistics** dialogue box.
 - Enter **Interval** in the **Variables** field.
 - Make sure that the **By Variables (optional)** field is empty.
 - Make sure that the Minitab output will include the mean and standard deviation by selecting these two summary statistics after clicking the **Statistics...** button.

**Stat > Basic Statistics >
Display Descriptive
Statistics...**

The values of the mean and standard deviation given in the Minitab output are 330.1 and 254.7, respectively. As observed in Unit 5, these values are not too different, as you would expect if the data are observations from an exponential distribution.

Alternatively, you can find the mean and standard deviation using **Store Descriptive Statistics...**. In this case, the values given by Minitab are 330.149 and 254.745, respectively.

- (b) In the **Histogram: Simple** dialogue box, enter **Interval** in the **Graph variables** field. The scale of the resulting histogram can be edited by obtaining the **Edit Bars** dialogue box (by selecting the bars and double-clicking on them) and, on the **Binning** tab, selecting **Cutpoint** and **Midpoint/Cutpoint positions**, then entering **0:1200/100** in the **Midpoint/Cutpoint positions** field. Finally, delete the title. The resulting histogram is shown in Figure S11.

Recall that the mean and standard deviation of an exponential distribution are equal.

Graph > Histogram... and select **Simple**.

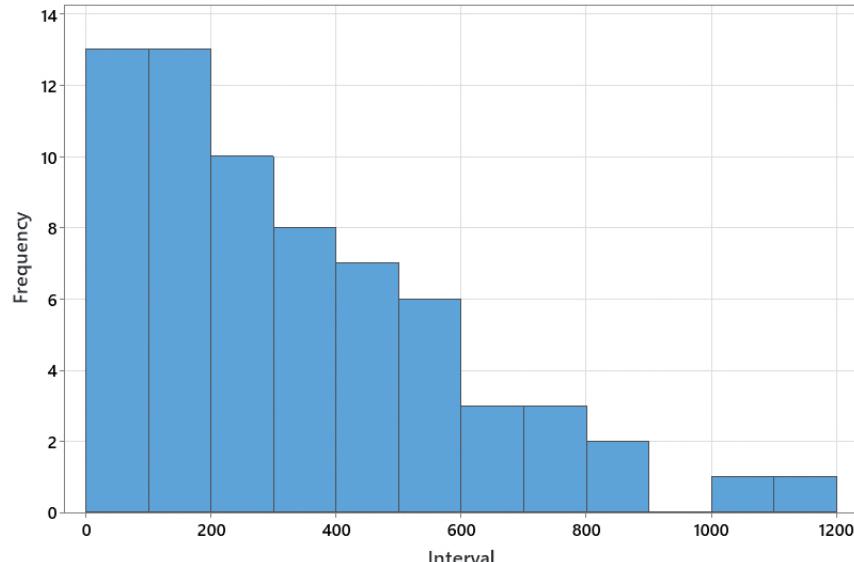


Figure S11 A histogram of the waiting times between serious earthquakes

As observed in Unit 5, the highest frequencies are for the shortest intervals. The frequencies tend to tail off for higher waiting times. This general shape is consistent with the data being observations from an exponential distribution.

Solution to Computer activity 38

- (a) Minitab gives the values 213.4 and 313.5 for the sample mean and sample standard deviation. A histogram of the data is shown in Figure S12.

The scale was edited to produce this histogram by selecting **Cutpoint under Interval**. Type in the **Binning** panel of the **Edit Bars** dialogue box and entering $0:2400/100$ in the **Midpoint/Cutpoint** positions field.

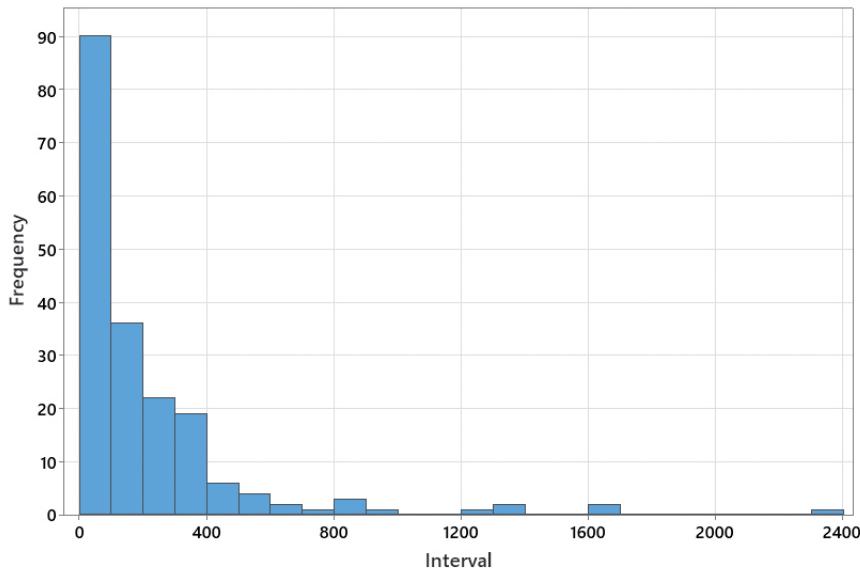


Figure S12 Intervals between coal-mining explosions

The histogram is highly skewed with a peak for low values, as expected for data from an exponential distribution. However, the values of the mean and standard deviation are not very close. There is perhaps some doubt about whether the waiting times between explosions are exponentially distributed.

- (b) A scatterplot showing the number of explosions that have occurred against time is shown in Figure S13.

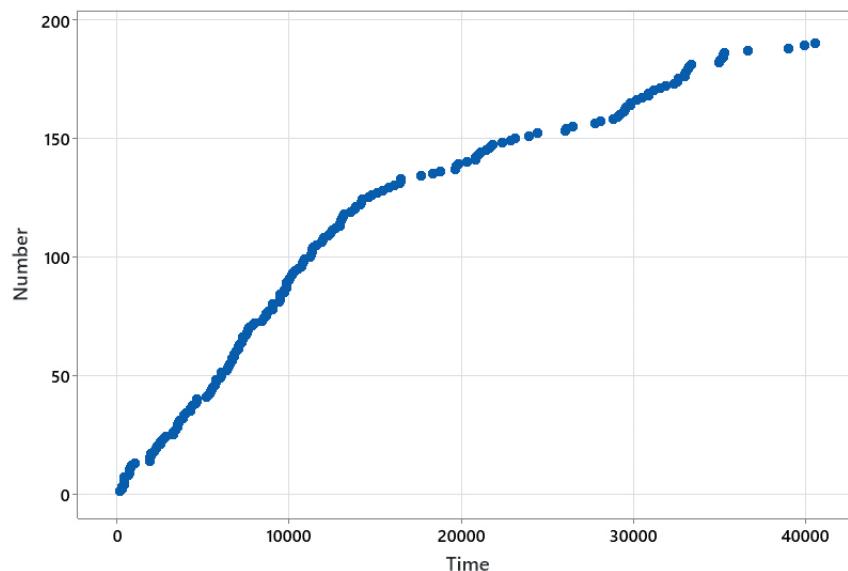


Figure S13 A scatterplot of explosions against time

It looks as though the rate of occurrence of coal-mining explosions decreased during the period of observation. Explosions were less frequent towards the end of the period. This is consistent with safety in mines improving from the nineteenth to the twentieth centuries.

- (c) The rate of occurrence of coal-mining explosions does not appear to have remained constant over the period of observation. So a Poisson process is not a suitable model for the occurrences of coal-mining explosions. There is also a little doubt about whether an exponential distribution is a good model for the waiting times between explosions.

Solution to Computer activity 39

The scatterplot is shown in Figure S14.

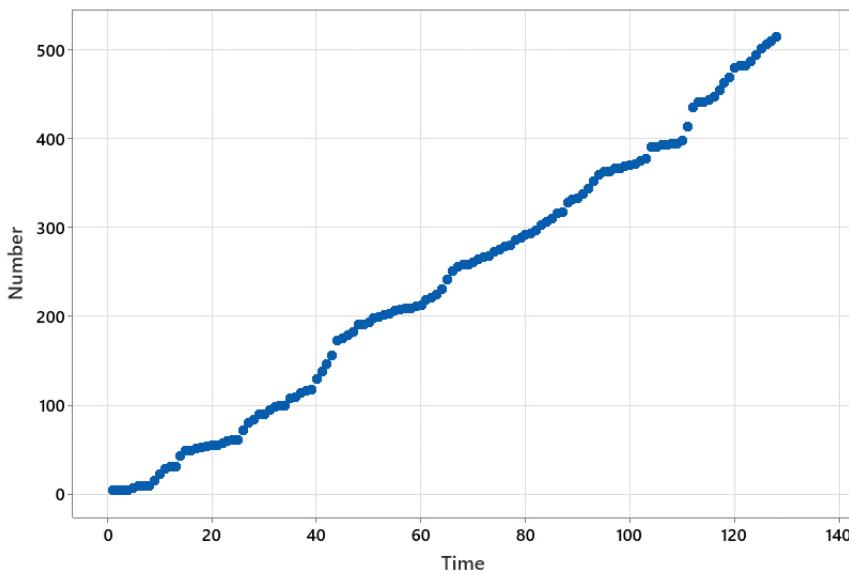


Figure S14 Cumulative requests to review papers against time

The points lie roughly along a straight line, so the rate of occurrence of requests to review papers appears to have remained fairly constant over the period of observation.

Solution to Computer activity 40

Stat > Basic Statistics >
Display Descriptive
Statistics...

- (a) Minitab gives the values 4.016 and 3.808 for the sample mean and sample standard deviation. So the sample variance is $(3.808)^2$ or approximately 14.5. The mean and variance of a Poisson distribution are equal. However, the values of the sample mean and sample variance are not close. These results suggest that a Poisson distribution is unlikely to be a good model for the number of requests received in a week.

(b) The histogram is shown in Figure S15.

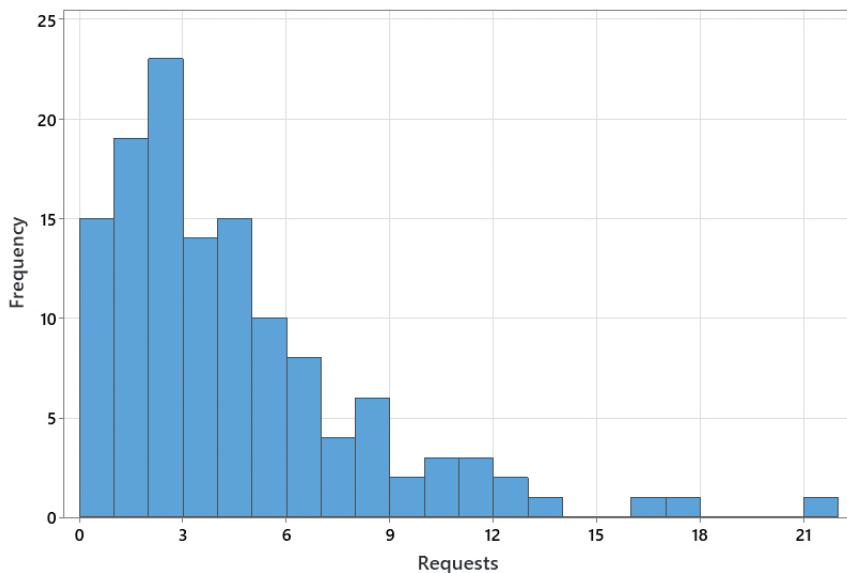


Figure S15 A histogram of the review requests data

(c) The diagram required is shown in Figure S16.

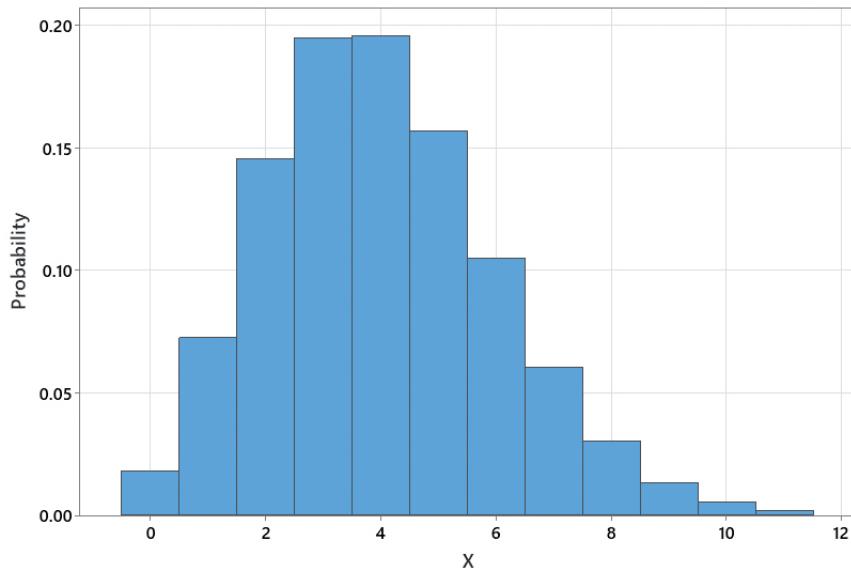


Figure S16 The probability mass function of a $\text{Poisson}(4.016)$ distribution

Notice that, although the Poisson distribution is discrete, there are no spaces between bars in the probability mass function produced by Minitab, so that the probability mass function produced by Minitab looks like a histogram.

Although both diagrams are right-skew, the histogram of the data is more strongly skewed than the probability mass function: its peak is further to the left than the peak of the p.m.f. The Poisson distribution with mean 4.016 is not a good model for the number of requests in a week. So a Poisson process is unlikely to be a good model for the occurrences of requests for review of papers in this department.

Solution to Computer activity 41

Allowing for a leap year, there are 1461 days in a four-year period.

Calc > Probability Distributions > Poisson...

- (a) If X is a random variable representing the number of serious earthquakes that occur in a typical four-year period, then X has a Poisson distribution with parameter

$$\lambda t = \frac{67}{22\,120} \times 1461 \simeq 4.425.$$

- (b) Obtain the **Poisson Distribution** dialogue box. Enter the value 4.425 in the **Mean** field. You need to select **Probability** for the first probability and **Cumulative probability** for the other two. In each case, you must enter an appropriate value in the **Input constant** field. The first values obtained with six decimal places in each part of the solution below are those given by Minitab.

- (i) The probability required is

$$P(X = 5) = 0.169290 \simeq 0.1693.$$

- (ii) The probability required is

$$P(X < 3) = P(X \leq 2) = F(2) = 0.182191 \simeq 0.1822.$$

- (iii) The probability required is

$$\begin{aligned} P(X \geq 8) &= 1 - P(X \leq 7) = 1 - F(7) \\ &= 1 - 0.919462 = 0.080538 \simeq 0.0805. \end{aligned}$$

Recall that Minitab does not provide the facility to work out $P(X \geq 8)$ directly: you have to work out $P(X \leq 7)$ and then calculate $P(X \geq 8) = 1 - P(X \leq 7)$. In particular, the **Inverse cumulative probability** field *does not* result in this value. (It will be used for its proper purpose in Chapter 12.)

- (c) If T is a random variable representing the waiting time in days between successive serious earthquakes, then T has an exponential distribution with parameter $\lambda = 67/22\,120$.

The mean waiting time in days is $1/\lambda = 330$ (to the nearest day).

- (d) Obtain the **Exponential Distribution** dialogue box. Enter 330 in the **Scale** field and 0 in the **Threshold** field of the **Exponential Distribution** dialogue box. You need to select **Cumulative probability** for both calculations.

Calc > Probability Distributions > Exponential...

- (i) The probability required is

$$P(T < 30) = F(30) = 0.0868993 \simeq 0.0869.$$

- (ii) Assuming that neither year is a leap year, there are 730 days in a two-year period, so the probability required is

$$\begin{aligned} P(T > 730) &= 1 - P(T \leq 730) = 1 - F(730) \\ &= 1 - 0.890532 \simeq 0.1095. \end{aligned}$$

Remember that T is a continuous random variable, so $P(T < 30) = P(T \leq 30) = F(30)$.

Solution to Computer activity 42

- (a) If X is a random variable representing the number of alpha particles emitted in a one-minute period, then X has a Poisson distribution with parameter

$$\lambda t = 0.517 \times 60 = 31.02.$$

- (b) Obtain the **Poisson Distribution** dialogue box. Enter 31.02 in the **Mean** field. Select **Probability** for the first probability and **Cumulative probability** for the other two. In each case you must enter an appropriate value in the **Input constant** field.

Calc > Probability Distributions > Poisson...

- (i) The probability required is

$$P(X = 30) = 0.0714133 \simeq 0.0714.$$

- (ii) The probability required is

$$P(X < 20) = P(X \leq 19) = F(19) = 0.0142899 \simeq 0.0143.$$

- (iii) The probability required is

$$\begin{aligned} P(X \geq 50) &= 1 - P(X \leq 49) = 1 - F(49) \\ &= 1 - 0.998959 \simeq 0.0010. \end{aligned}$$

- (c) If T is a random variable representing the intervals between successive emissions, then T has an exponential distribution with parameter $\lambda = 0.517$.

Calc > Probability Distributions > Exponential...

- (d) Obtain the **Exponential Distribution** dialogue box.

Since $\lambda = 0.517$, the mean of the distribution is $1/\lambda \simeq 1.934$, so enter 1.934 in the **Scale** field (and 0 in the **Threshold** field). You need to select **Cumulative probability** for both calculations.

- (i) The probability required is

$$P(T < 0.5) = F(0.5) = 0.227815 \simeq 0.2278.$$

- (ii) The probability required is

$$P(T > 3) = 1 - F(3) = 1 - 0.788004 \simeq 0.2120.$$

Solution to Computer activity 43

Initially, the value of α on the slider bar is 0.5, so q_α is initially $q_{0.5}$, which is the median.

As the value of α decreases, so the value of q_α decreases: the value of α moves down the vertical axis of the c.d.f. so that q_α moves to the left, while on the p.d.f. the shaded area decreases in size which also moves q_α to the left. As the value of α increases, the opposite happens and the value of q_α increases: the value of α moves up the vertical axis of the c.d.f. so that q_α moves to the right, while on the p.d.f. the shaded area increases in size moving q_α to the right.

Solution to Computer activity 45

- (a) In the **Exponential Distribution** dialogue box, since $q_L = q_{0.25}$, enter 0.25 in the **Input constant** field.

The Minitab output states that $P(X \leq 94.9351) = 0.25$, so the lower quartile is approximately 95 days.

Similarly, $q_U = q_{0.75}$, so input the value 0.75 in the **Input constant** field. From the Minitab output, the upper quartile is approximately 457 days.

Hence the interquartile range is approximately $457 - 95 = 362$ days. (Minitab does not have a facility for calculating the interquartile range of a distribution directly.)

- (b) Either type the values 0.1, 0.2, ..., 0.9 directly in column C1 of the worksheet, or use **Calc > Make Patterned Data** as follows. In the **Simple Set of Numbers** dialogue box, enter C1 in the **Store patterned data in** field, 0.1 in the **From first value** field, 0.9 in the **To last value** field, and 0.1 in the **In steps of** field.

When finding the deciles, select **Input column** in the **Exponential Distribution** dialogue box and enter C1 in its field, and enter C2 in the **Optional storage** field. The deciles, rounded to the nearest integer, are listed in the table below.

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
q_α	35	74	118	169	229	302	397	531	760

- (c) Since $q_{0.1} \approx 35$, approximately 10% of intervals between serious earthquakes are shorter than 35 days.

Since $q_{0.8} \approx 531$, approximately 80% of intervals between serious earthquakes are shorter than 531 days and therefore approximately 20% exceed 531 days.

Calc > Make Patterned Data > Simple Set of Numbers...

Calc > Probability Distributions > Exponential...

Solution to Computer activity 46

- (a) In the **Binomial Distribution** dialogue box, if the value 0.25 is entered in the **Input constant** field, then Minitab displays the following results:

$$P(X \leq 2) = 0.206085, \quad P(X \leq 3) = 0.411449.$$

So the lower quartile q_L is 3.

Similarly, entering the value 0.75 in the **Input constant** field leads to

$$P(X \leq 4) = 0.629648, \quad P(X \leq 5) = 0.804208.$$

So the upper quartile is 5.

- (b) At least 999 out of 1000 students who guess must score less than the pass mark, so the pass mark x must be chosen so that $P(X < x) \geq 0.999$, that is, $P(X \leq x - 1) \geq 0.999$. So find $q_{0.999}$; this will give the highest score that will be a fail.

According to Minitab, $P(X \leq 9) = 0.997405$ (which is less than 0.999) and $P(X \leq 10) = 0.999437$ (which exceeds 0.999). So $q_{0.999}$ is equal to 10.

So if the highest score that will fail the exam is 10, then the pass mark should be 11.

For a discrete random variable, $P(X < x) = P(X \leq x - 1)$.

Solution to Computer activity 47

- (a) For a Poisson distribution with mean 4.425, in the **Poisson Distribution** dialogue box when you enter the value 0.5 in the **Input constant** field, Minitab gives the two probabilities

$$P(X \leq 3) = 0.355107, \quad P(X \leq 4) = 0.546396.$$

So according to the Poisson model, the median number of serious earthquakes in a four-year period is 4.

Calc > Probability Distribution > Poisson...

- (b) (i) For a Poisson distribution with mean 40, in the **Poisson Distribution** dialogue box when you enter the value 0.1 in the **Input constant** field, Minitab gives

$$P(X \leq 31) = 0.0855206, \quad P(X \leq 32) = 0.115304.$$

So the largest number of calls received during any of the 10% of hours that are the quietest is 32.

- (ii) When you enter 0.01 in the **Input constant** field, you obtain

$$P(X \leq 25) = 0.0075664, \quad P(X \leq 26) = 0.0123106.$$

So even during one of the 1% of hours that are the quietest, as many as 26 calls may be received.

Solutions to computer exercises

Solution to Computer exercise 1

- (a) The required histogram is shown in Figure S17.

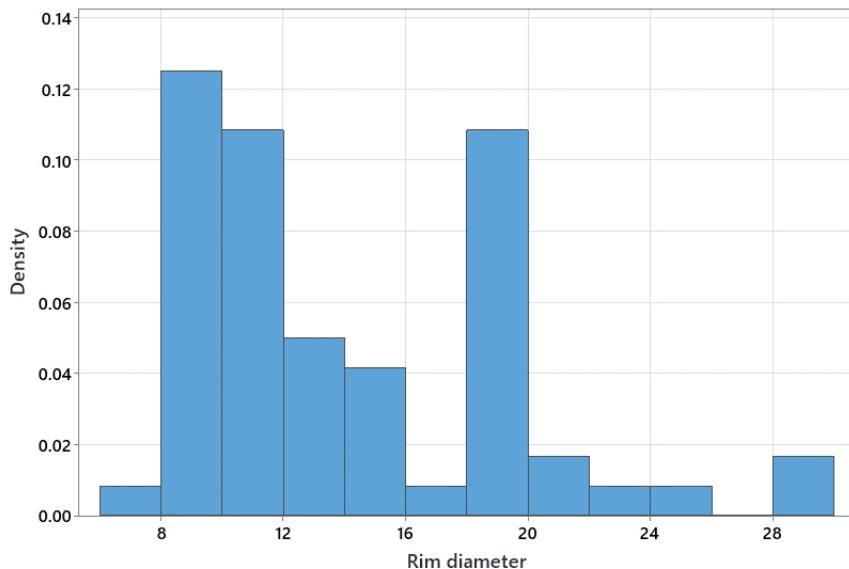


Figure S17 Unit-area histogram of rim diameters

The histogram was obtained by selecting **Graph > Histogram...** and then **Simple**, entering **Rim diameter** in the **Graph variables** field of the **Histogram: Simple** dialogue box, and in the **Histogram: Scale** dialogue box (obtained via the **Scale...** button), ensuring that the **Y-Scale Type** option on the **Y-Scale Type** tab is **Density**. After producing the default unit-area histogram, the bins were changed by obtaining the **Edit Bars** dialogue box (by selecting and then double-clicking on a bar) and on the **Binning** tab, selecting **Cutpoint** and entering **6:30/2** in the **Midpoint/Cutpoint positions** field.

The histogram shows an essentially bimodal structure to these data, showing two main groups of cups, many with rim diameters around 10 cm and fewer with rim diameters around 19 cm. (A small number of cups are much larger, around 29 cm in diameter.)

- (b) The required scatterplot is shown in Figure S18.

Graph > Scatterplot... and select **Simple**.

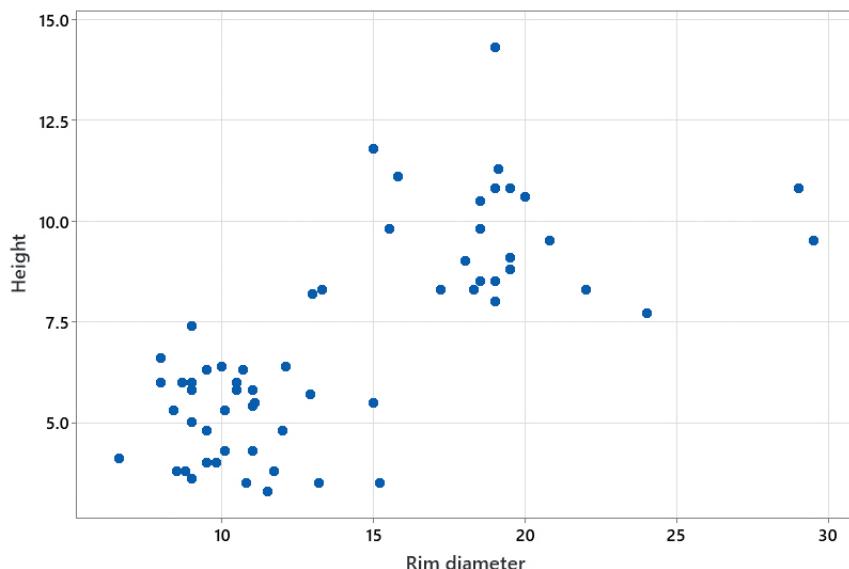


Figure S18 Scatterplot of height against rim diameter

The scatterplot shows a positive relationship between height and rim diameter. The relationship seems quite linear except perhaps for either a ‘curving down’ or else some possible outliers towards the right-hand side of the plot. The relationship between height and rim diameter does not seem especially strong because there appears to be quite a lot of scatter about any trend line.

Something else that you probably noticed in the scatterplot in Figure S18 is that these Bronze Age cups divide into two ‘clusters’, the small (with small heights and small rim diameters) and the large (with large heights and large rim diameters), plus, perhaps, a small number of cups that don’t fit this characterisation so well.

- (c) Given that the distribution of rim diameters is bimodal and that there is a positive relationship between height and rim diameter, you might expect the distribution of cup heights to be bimodal too. In fact, the histogram of cup heights in Figure S19 supports the notion that the distribution of heights is bimodal.

‘Cluster analysis’ is itself an important statistical technique, but not one that will be covered in this module.

(The results of all parts of this question are related to the strong suggestion directly from the scatterplot in Figure S18 that these Bronze Age cups divide into two clusters, as described in the solution to part (b) above.)

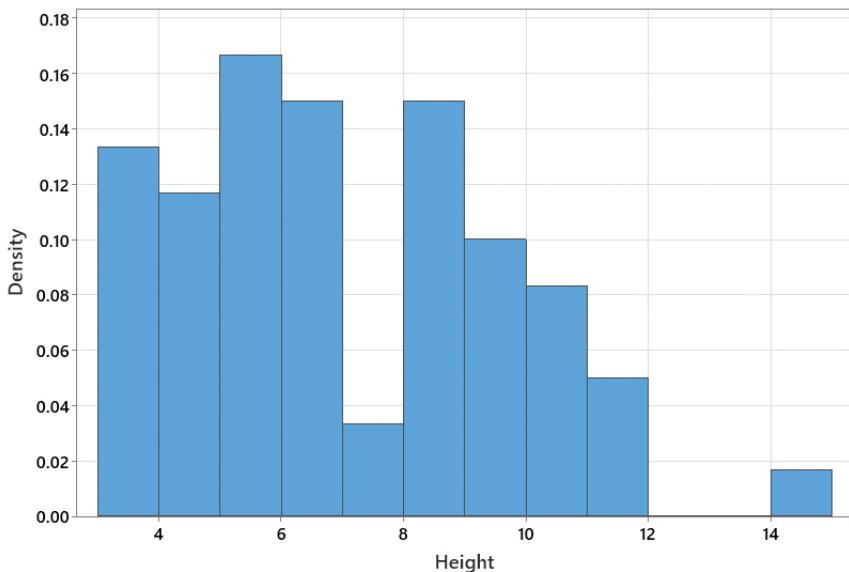


Figure S19 Unit-area histogram of cup heights

Solution to Computer exercise 2

- (a) On the worksheet, the data for the number of employees are given in a single column, with another column being used to give the value of the industry code. The comparative boxplot is therefore obtained by selecting **Graph > Boxplots... ,** and then **With Groups** under **One Y.**

In the **Boxplot: One Y, With Groups** dialogue box, enter **Employees** in the **Graph variables** field and **Industry code** in the **Categorical variables for grouping (1-4, outermost first)** field. To produce horizontal boxplots, obtain the **Boxplot: Scale** dialogue box (by clicking on the **Scale...** button), and ensure that the **Transpose value and category scales** option is selected.

The resulting comparative boxplot is shown in Figure S20.

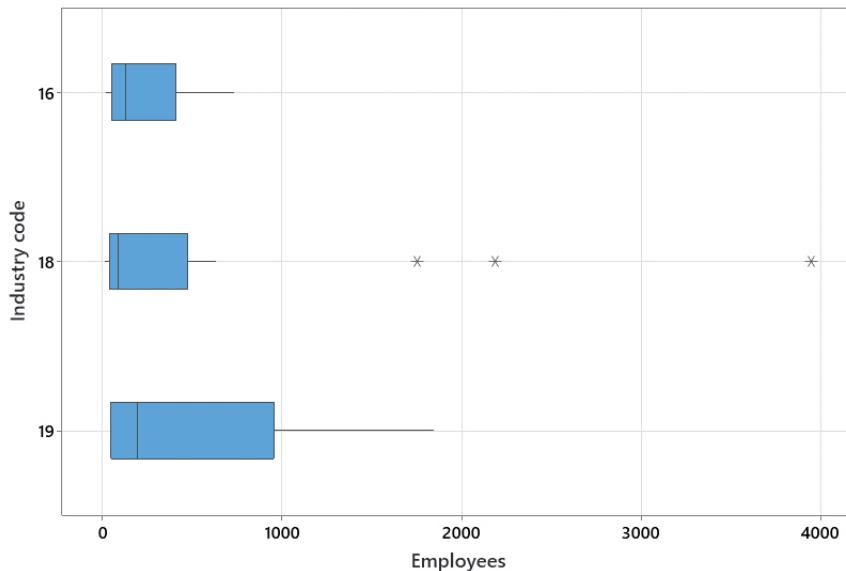


Figure S20 Numbers of employees in each industry

- (b) The boxplots indicate that the numbers of employees in companies with industry codes ‘16’ and ‘18’ have similar distributions apart from the three much larger companies in industry ‘18’. All distributions are right-skew; the boxplot and interquartile range for industry group ‘19’ also indicate considerable extra spread in numbers of employees compared with that in the other two industries.

Solution to Computer exercise 3

- (a) Assuming that all passengers turn up (or not) independently of each other, it is reasonable to model the number of passengers who do not turn up by a binomial distribution: $X \sim B(140, 0.1)$.
- (b) The number of standby passengers who get seats on the flight is equal to the number of passengers out of the 140 with reservations who do not turn up for the flight. All the standby passengers will get seats on the flight if at least 16 of the 140 passengers who have reserved seats do not turn up. So the probability required is

$$P(X \geq 16) = 1 - P(X \leq 15).$$

This probability may be found using **Calc > Probability Distributions > Binomial...** In the **Binomial Distribution** dialogue box, select **Cumulative probability**, enter 140 and 0.1 for the parameters of the distribution, and enter 15 in the **Input constant** field. Then

$$P(X \geq 16) = 1 - P(X \leq 15) = 1 - 0.674802 = 0.325198 \simeq 0.325.$$

- (c) The probability that exactly half the standby passengers get seats on the flight is given by $P(X = 8)$. Select **Probability** in the **Binomial Distribution** dialogue box, and enter 8 in the **Input constant** field. You will find that

$$P(X = 8) = 0.0272309 \simeq 0.027.$$

Solution to Computer exercise 4

- (a) If the waiting times are observations from an exponential distribution, then their mean and standard deviation should be approximately equal. Using **Stat > Basic Statistics > Display Descriptive Statistics...**, the sample mean and standard deviation are 50.60 hours and 43.30 hours, respectively. These are quite similar, so an exponential model could be plausible.

A histogram of the data is shown in Figure S21.

Graph > Histogram... and select Simple.

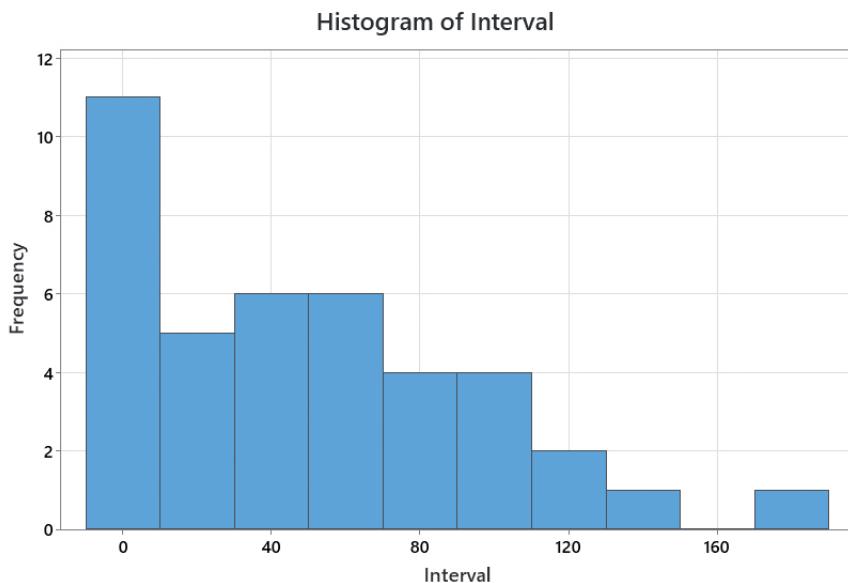


Figure S21 Intervals between admissions at an intensive-care unit

The shape of the histogram is not inconsistent with the data coming from an exponential distribution. So, since the mean and standard deviation are also quite close, it looks like an exponential model may be suitable.

- (b) The function **Partial sum**, available when using **Calc > Calculator...**, was used to obtain the partial sums of the waiting times in a column called **Time**. **Calc > Make Patterned Data > Simple Set of Numbers...** was used to enter the numbers 1, 2, ..., 40 in a column called **Number**.

See Computer activity 36 for fuller instructions.

Graph > Scatterplot... and select Simple.

Figure S22 shows a scatterplot of the number of admissions against time.

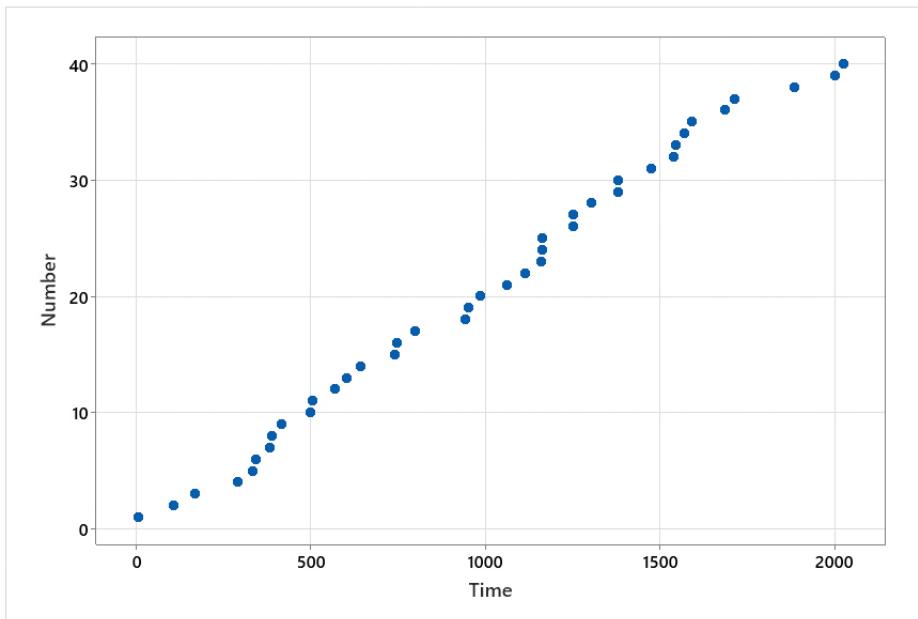


Figure S22 A scatterplot for the admissions data

The points lie roughly along a straight line through the origin, suggesting that the average rate of admissions to the intensive-care unit remained constant over the period to which the data relate.

In fact, there are variations in the underlying rate of admission, both with the time of day and with the day of the week, which can be detected when the full set of the original data is analysed.

Solution to Computer exercise 5

- (a) If X is a random variable representing the number of major explosive volcanic eruptions that occur in a typical ten-year ($= 10 \times 12 = 120$ -month) period, then X has a Poisson distribution with parameter

$$\lambda t = 0.0352 \times 120 = 4.224.$$

The probability required is

$$P(X > 5) = 1 - P(X \leq 5) = 1 - F(5).$$

Choose **Calc > Probability Distributions > Poisson...**. Enter the value 4.224 for the **Mean** in the **Poisson Distribution** dialogue box. Select **Cumulative probability** and enter 5 in the **Input constant** field. Minitab returns the value 0.749214, so the probability required is

$$P(X > 5) = 1 - 0.749214 \simeq 0.2508.$$

- (b) If T is a random variable representing the interval in months between successive major explosive volcanic eruptions, then T has an exponential distribution with parameter $\lambda = 0.0352$.

The mean interval between eruptions (in months) is $1/\lambda = 1/0.0352 \simeq 28.4$.

The probability required is $P(T < 6)$. Choose **Calc > Probability Distributions > Exponential...** and enter the value 28.4 in the **Scale** field in the **Exponential Distribution** dialogue box. Select **Cumulative probability** and enter 6 in the **Input constant** field.

According to Minitab, the proportion of intervals that will be less than six months is

$$P(T < 6) = P(T \leq 6) = F(6) = 0.190443 \simeq 0.1904.$$

- (c) Since 5% of intervals are shorter than x months, x is the 0.05-quantile of the exponential distribution with mean 28.4; that is, $x = q_{0.05}$.

Choose **Calc > Probability Distributions > Exponential...** and enter the value 28.4 in the **Scale** field in the **Exponential Distribution** dialogue box. Select **Inverse cumulative probability** and enter 0.05 in the **Input constant** field.

The Minitab output gives $x = 1.45673$. So only 5% of intervals are shorter than about $1\frac{1}{2}$ months.

- (d) Since only 1% of intervals exceed y years, or $12y$ months, $12y$ is the 0.99-quantile of the exponential distribution with mean 28.4; that is, $12y = q_{0.99}$.

Proceed as in part (c), but enter 0.99 in the **Input constant** field.

The Minitab output gives the value 130.787, so $12y = 130.787$. Hence

$$y = \frac{130.787}{12} \simeq 10.90.$$

So only 1% of intervals are longer than about 11 years.

Index

- American weights animation, 36
- Bar Charts**, 10
- Basic Statistics**, 23
- Binning**, 18
- binomial approximation of a Poisson distribution, 46
- binomial probability, 39
- Boxplots**, 20
- Calculator**, 52
- c.d.f., 41
- comparative boxplot, 29
- discrete uniform distribution, 44
- Display Descriptive Statistics**, 23
- editing a graph, 12
- exponential distribution, 53, 57
- Histograms**, 17
- horizontal boxplot, 21
- Make Patterned Data**, 42
- new worksheet, 42
- numerical summary, 23
- opening a worksheet, 7
- partial sum, 52
- pasting output into a word-processor document, 16
- p.m.f., 40
- Poisson distribution, 48, 56
- Poisson process, 51
- printing output, 16
- Probability Distribution Plots**, 56
- Probability Distributions**, 39
- project file, 14
- pseudo-random number, 39
- quantile, 59, 61
- Quantiles** animation, 59
- raw form, 10
- Royal deaths** animation, 44
- running Minitab, 5
- saving your work, 14, 15
- Scatterplot**, 33
- Score on a die** animation, 34
- side-by-side bar charts, 26
- simulation, 34
- Store Descriptive Statistics**, 24
- summary form, 13
- unit-area histogram, 28
- V-1 bombs** animation, 48
- worksheet, 7
- Worksheet Description**, 8