UNIT 2 GRAPHING DATA

AIMS

To illustrate the most appropriate use of tables and graphs for displaying data.

OBJECTIVES

At the end of the second week you should be able to:

- Choose and draw the most appropriate graph to present different types of data, distinguishing between bar and pie charts, boxplots, dotplots, histograms, step charts and ogives (cumulative frequency graphs).

- Interpret such graphs.

Reading: Bland, sections 5.3, 5.4, 5.5, 5.8.
or Bowers, pp. 49-57

Introduction

If we want to display data graphically, we need first to identify the type of data involved (see Unit 1), because some charts are appropriate with some types of data but not with other types. We will discuss a number of different charts, and indicate, with each type of chart, which type of data it is appropriate for.

Bar chart

A bar chart can be used to display either *nominal*, *ordinal* or *discrete metric* values. The bar chart is drawn as a set of vertical columns, one for each category or value, the heights of which are equal to their respective frequencies. The bars are of equal but arbitrary width; spaces between the bars emphasise the non-continuous nature of the data. The number of values which can be plotted is limited only by the width of the page, the need for clarity in the figure, etc. One variable (**simple** bar chart), or several variables (either **clustered** or **stacked** bar charts) may be plotted. Figure 2.1 shows a simple bar chart for the four categories of hair colour (data in Table 1.1, Unit 1).

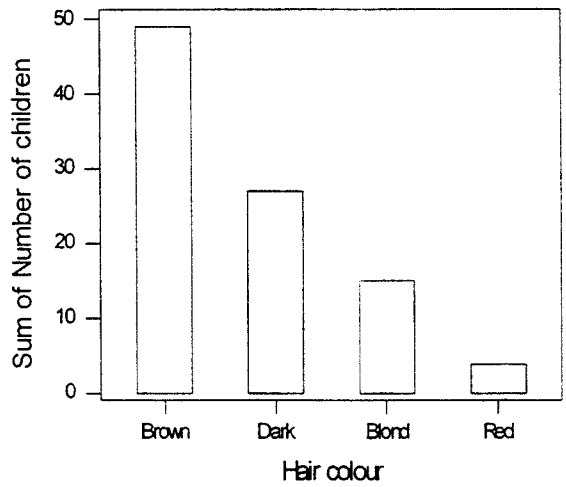


Figure 2.1 Simple bar chart of hair colour data (see Table 1.1)

With nominal data, unless there is a particular reason why not, it's good practice, as here, to arrange the bars in descending or ascending order of height (Brown, Dark, Blond and Red in this example). With ordinal or discrete metric data, the bars should be in the same order as the categories or values.

Figure 2.2 shows a simple bar chart displaying scores by a sample of 5362 men on the Psychiatric Symptom Frequency Scale (or PSF)*. This scale has possible scores from 0 to 100 (high scores indicate more severe the symptoms), but the scores charted are only for those men scoring between 0 and 35.. Notice that the majority of men had a PSF score of 10 or less. This chart must be getting somewhere near the practical maximum number of categories (values) that can be usefully displayed.

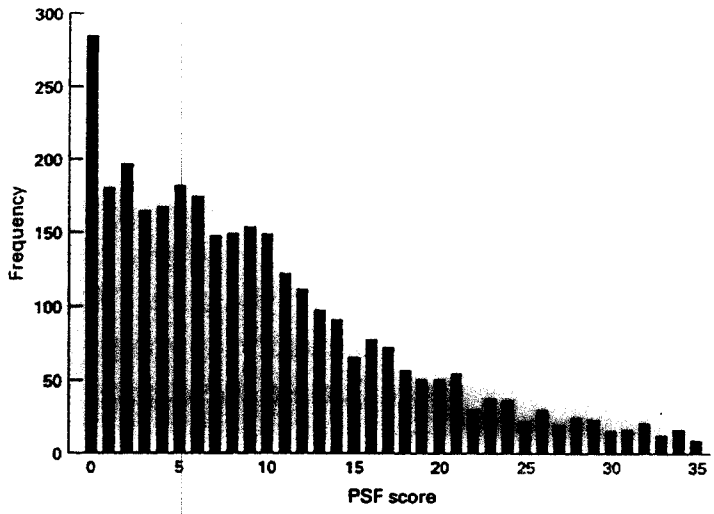


Figure 1 Frequency distribution of the lowest 95% of the psychiatric symptom frequency (PSF) score.

Figure 2.2 Simple bar chart showing ordinal scores on the PSF for a sample of 5362 men. J Epid & Community Health, 51, 1997

Q. 2.1 How would you describe the shape of the distribution of PSF scores in Figure 2.2?

The **clustered** bar chart is most appropriate when there is a need to compare the size of sub-groups *within* each category. The **stacked** bar chart is best used to compare the *total* sizes *across* the categories, since sub-group comparisons can be difficult if the relative sizes of sub-groups change radically between categories (see below for example of a stacked bar chart).

Figure 2.3 is a clustered bar chart showing the distribution of symptom duration (months) in four groups of patients suffering from lateral epicondylitis (tennis elbow). Each group received a different treatment - A or B or C or D.

* This scale was developed to assess symptoms of anxiety and depression in the general population.

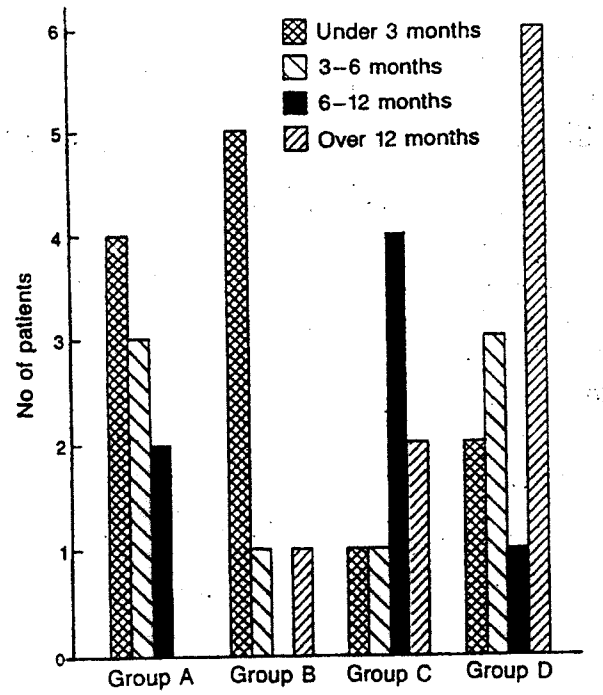


Fig 1: Distribution of symptom duration within the groups

Figure 2.3 Clustered bar chart - duration of tennis elbow by treatment group

Its easy to compare the number of patients in each of the four symptom-duration sub-groups within each treatment category. For example, in treatment group C, those with symptoms lasting 6-12 months formed the largest sub-group (4 patients), those with symptoms lasting under 3 months and 3-6 months equally the smallest (1 patient each).

Q. 2.2 In Figure 2.3, which duration of symptom sub-group in treatment Group A had: (a) most patients? (b) the fewest?

Figure 2.4 is a **stacked** bar chart (lying on its side) which shows the results of a patient satisfaction questionnaire (the Patient Intentions Questionnaire or PIQ) among 504 patients attending 25 different GPs. The chart records the gaps between what the patient wanted from the consultation and what was actually achieved. The chart makes it easy to compare the % *total* number of patients in each "want" category.

Q. 2.3 In Figure 2.4 which "Want" had: (a) the smallest total % of patients; (b) the largest; (c) the biggest gap between want and satisfaction of that want?

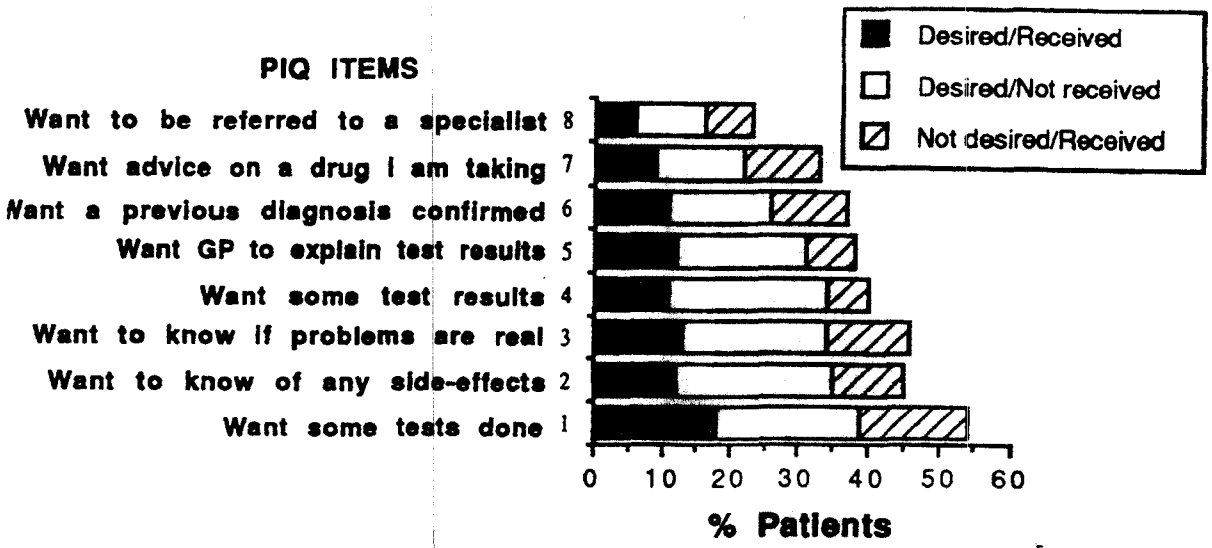


FIGURE 4 *Proportion of patients desiring and receiving 'test and diagnosis' items from the Patients' Intentions Questionnaire (PIQ) (n = 420)*

Figure 2.4 Stacked bar chart showing the % patients with wants in eight different areas. Source:

Pie Chart

A pie chart may be used for both *nominal* and *ordinal* data. It works best if there are only two or three categories. The chart is drawn as a circle divided into segments, one per category or value, the angle of each segment proportional to the frequency of that category. Only *one* variable may be plotted with each pie chart. Figure 2.5 is a pie chart for the hair colour data (Table 1.1).

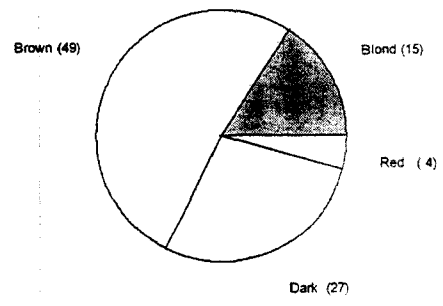


Figure 2.5 Pie chart of hair colour (data in Table 1.1)

Q. 2.4 Sketch bar and pie charts for the data shown in Table 2.1 which shows

the ethnic origin of subjects in a sample of 51 schizophrenics patients recruited for a study into their use of mental health services and interviewed two years after discharge from hospital. What advantages and disadvantages do these charts have over the original frequency table on its own?

Ethnic origin	Number of patients (n=51)
Afro-Caribbean	20
Asia	3
White	27
Other	1

Table 2.1 Frequency table for the ethnic origin of 51 schizophrenic patients. BMJ, 308, 1994

The dotplot

A dotplot can best be used to display either discrete metric or ordinal data, although might not be very useful with metric continuous data. Each sample value is represented by a dot on a graph with frequency on the vertical axis and value on the horizontal. The dotplot is excellent for showing how the sample values are distributed, and what shape the distribution has.

Figure 2.6 shows the dotplot of the ages of a sample of 1711 deliberate self-harm patients referred to a psychiatric service in a large general hospital. Notice that although age is a continuous metric variable it has been recorded here as *if* discrete, i.e. as age last birthday. This makes it possible to use a dotplot.

Q. 2.5 In Figure 2.6 (a) Approximately what is the commonest age of these subjects, (b) How would you describe the shape of this distribution?

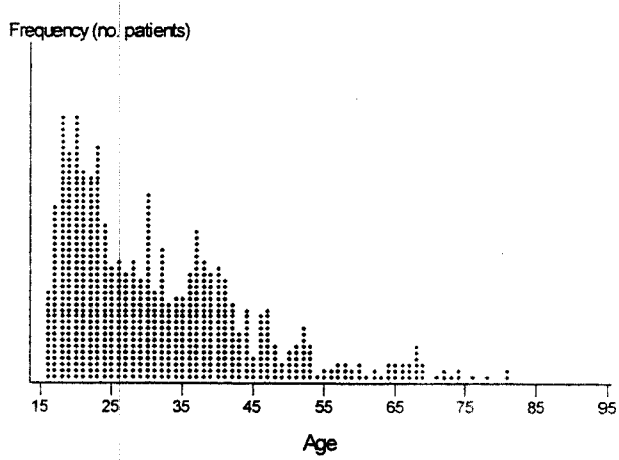


Figure 2.6 Dotplot of age for a sample of 1711 deliberate self-harm patients (unpublished data)

The histogram

A histogram is used to display continuous data when it has been *grouped*. As well as providing us with a view of the relative size of the frequencies in each group or class, it also enables us to get an impression of the shape of the distribution. Each column in the histogram represents the frequency of the corresponding group. Notice that there is no gap between adjacent columns emphasising that the data is continuous.

The area of each column (base × height) is proportional to that group's frequency and the total area of the histogram is proportional to total frequency. Figure 2.7 shows the frequency histogram for the data on the age distribution of the endometriosis women (Unit 1, Table 1.3, second column).

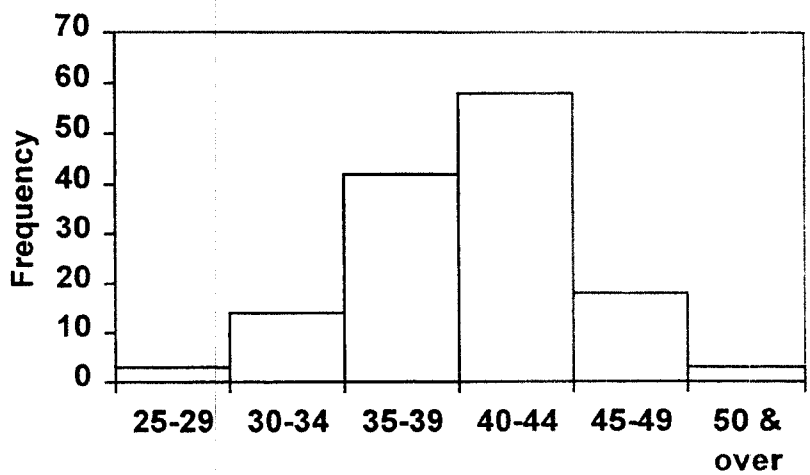


Figure 2.7 Histogram for the ages of endometriosis women (see Table 1.3)

Q. 2.6 What assumption has been made in drawing the histogram in Figure 2.7? Is this justified in this case?

Note that if all the groups have the same width then the height of each column is drawn equal to its respective frequency, as in Figure 2.7. But if some groups are of different widths, then the heights must be adjusted so that the areas are still proportional to the frequency. For instance consider the data in Table 2.2 which shows the longest length of stay in a mental hospital (in months) by each schizophrenic patient in the study referred to above (see Table 2.1).

Length of stay	Number of patients
0-3 months	17
3-6 months	15
6-12 months	11
12-24 months	8

Table 2.2 Length of stay in mental hospital by 51 psychiatric patients

Q. 2.7 What problem arises when you try to allocate into Table 2.2 data for a patient whose longest stay in hospital was 12 months? How accordingly would you modify the design of this frequency table?

Notice that the first and second groups (0-3 and 3-6 months) are each 3 months wide, but the third group (6-12 months) is 6 months wide - twice the width of the first two groups, and the last group is 12 months wide - four times as wide as the first two groups. If we use a width of 3 months as our "standard" width, then we can plot the heights of the first two groups at 17 and 15 respectively.

However, to preserve the area property, the height of the third group must be *halved* (since the width is *twice* the standard width) from 11 to 11/2 = 5.5. For the same reason, the height of the fourth group must be divided by 4 (since the group is four times the standard width). Table 2.3 shows the frequency table after appropriate adjustments to the frequencies. Any class can be chosen to give the standard width, but it helps calculations if the class width which is most numerous is used.

Length of stay	Adjusted number of patients
0-3 months	17
3-6 months	15
6-12 months	5.5
12-24 months	2

Table 2.3 Length of stay in mental hospital by 51 psychiatric patients with adjustment to frequencies in Table 2.2 to allow for differing group widths

Figures 2.8 and 2.9 show respectively histograms drawn incorrectly (making no allowance for differences in group widths) and after making appropriate adjustments to the frequencies of the last two groups (as in Table 2.3).

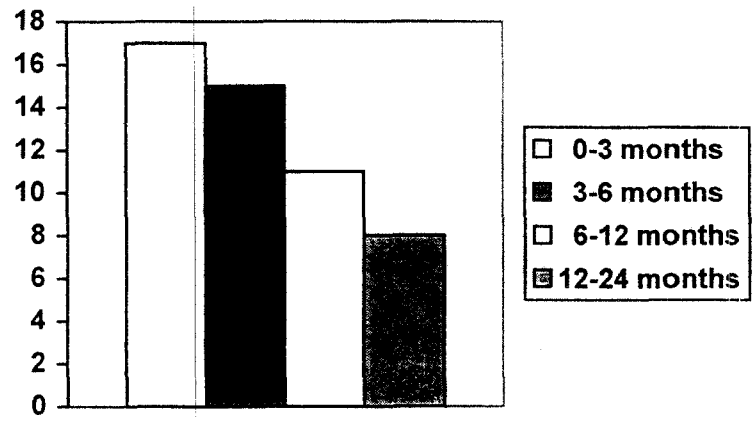


Figure 2.8 Incorrectly drawn histogram of the length of stay data in Table 2.3, making no allowance for the unequalness of the group widths

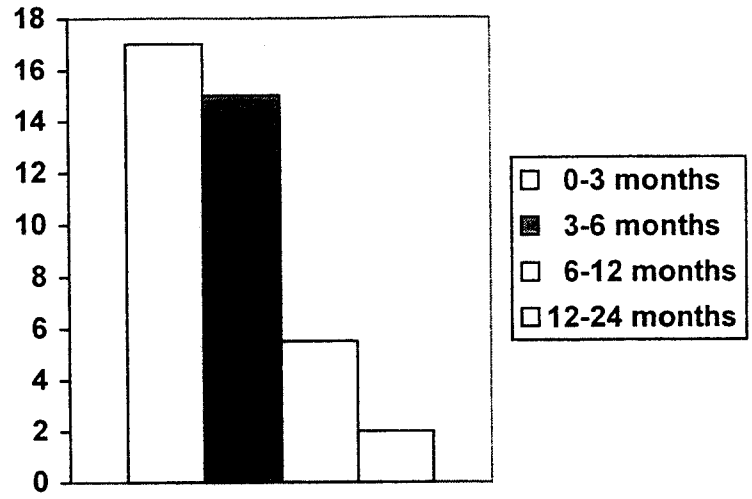


Figure 2.9 Correctly drawn histogram for length of stay data

Q. 2.8 What different messages are conveyed by the histograms in Figures 2.8 and 2.9?

Q. 2.9 Sketch a histogram using the grouped % frequency values you calculated in Q. 1.8 (Unit 1) on ICU % mortality rates. What can you say about the way % mortality is distributed?

The boxplot

The boxplot (also known as box-and-whisker plot) is a compact diagram useful for comparing the distributional features of a number of groups, and can be used with either ordinal or metric data. Figure 2.10 shows two boxplots from a study comparing the time taken by nurses to administer a traditional intramuscular injection (IMI) of analgesia for post-operative pain relief compared to the time taken to carry out safety checks on patients using patient-controlled analgesia (PCA).

The boxplot identifies (referring to Figure 2.10 from bottom to top):

(i) The minimum sample value. Usually shown at the bottom end of the "whisker" sticking down from the bottom of the box. Sometimes outliers* are used in place of the minimum.

* An outlier is a value much bigger or much smaller than the general mass of the values.

(ii) The value below which 25% (i.e. a quarter) of the values lie; known as the 25th percentile or the first quartile and denoted Q1. Shown as the bottom edge of the box.

(iii) The value below which 50% (i.e. half) of the values lie; known as the second quartile, Q2. Shown as the horizontal "shelf" near the middle of the box.

(iv) The value below which 75% (i.e. three quarters) of the values lie; known as the third quartile, Q3. Shown as the top edge of the box. Notice that the depth of the box (from Q1 to Q3) represents the range of the middle 50% of values.

(v) The maximum value. Usually shown at the top end of the "whisker" sticking up from the top of the box. Sometimes outliers are used in place of the maximum.

Figure 2.10 shows, for example, that the minimum time taken with the traditional method of pain relief was about six and a half minutes, compared to a minimum of about one and a half minutes with PCA. Three-quarters (75%) of the patients using the traditional method took less than about 12 and a half minutes, compared to about two and a half minutes with PCA.

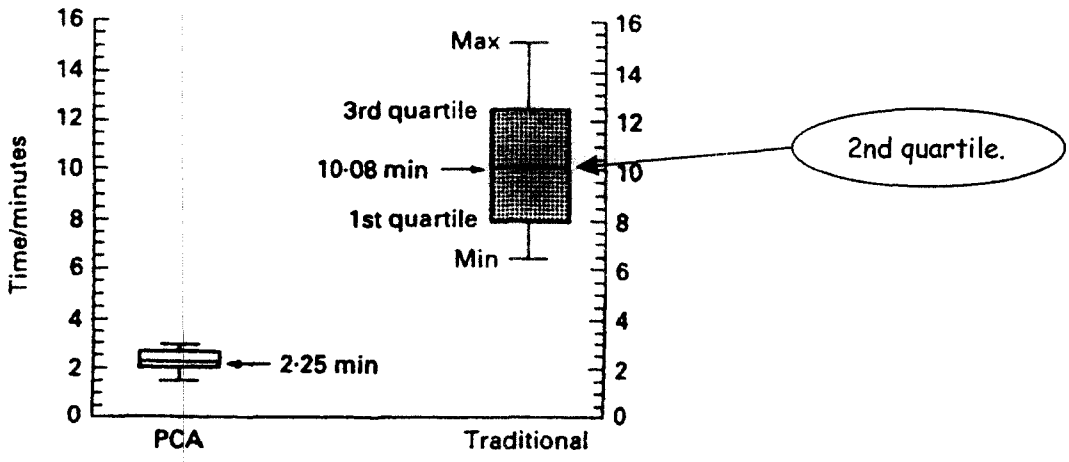


Figure 1 Box plot for time taken (arrow indicates the median time).

Figure 2.10. Boxplot showing time taken by nurses to give analgesia by injection (traditional method) with time taken to check patient-controlled analgesic (PCA). J of Advanced Nursing, 1994, 20.

Q. 2.10 (a) From Figure 2.10 how did maximum times of the two methods compare? (b) About half the patients took less than how long with the two methods? (c) What was the range in the time taken for the middle 50% of values using the traditional method?

Cumulative frequency curve (or ogive)

If we want to plot cumulative frequency we have two choices depending on data type. If the data is *continuous metric* we can use an **ogive** (which is a plot of cumulative frequency or relative cumulative frequency on the vertical axis, against value). If the data is either ordinal or discrete metric a **step** chart is appropriate (see below).

As an example, Figure 2.11 shows the relative cumulative frequency curves (or ogives) for the total cholesterol concentration (mmol/l) in both intervention and control subjects in a randomised control trial of the effectiveness of health checks conducted by nurses in a primary care setting. The intervention group received health advice from the nurses, the control group no intervention. Since this data is metric continuous an ogive is appropriate.

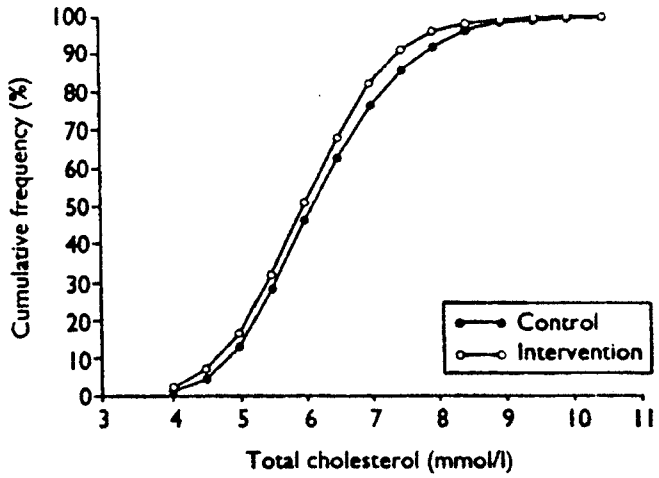


FIG 2—*Cumulative frequency distributions of total cholesterol concentration in control and intervention groups*

Figure 2.11 **Relative (%) cumulative frequency distribution (ogive) of total cholesterol concentration in control and intervention groups.**

Q. 2.11 From Figure 2.11, (a) approximately what percentage of patients in the control and intervention groups had cholesterol concentrations of less than 7mmol/l? (b) Half the patients in each group had cholesterol concentrations of less than what amount?

The step chart

If we wish to plot a cumulative frequency chart for ordinal data or for discrete metric data, the **step chart** is the most appropriate. Look again at the frequency distribution of the Disability Rating Scores in Table 1.2.

This data is ordinal and therefore discrete (ordinal data is necessarily discrete). Cumulative frequencies are: 1, 10, 12, 17, 22, 25, 25, 25, 27, and 28. A step chart of these cumulative frequencies is shown in Figure 2.12. The height of the step at any value is the corresponding cumulative frequency. For example, the height of the step at a DRS score of 4 is = 22, so 22 patients had a DRS score of 4 or less.

Steps charts are often used to chart survival and mortality data (where the horizontal axis will be time (days, weeks, etc.). The distinct step up from one value to the next (unlike the continuous ogive in Figure 2.11) emphasises that this data is discrete not continuous.

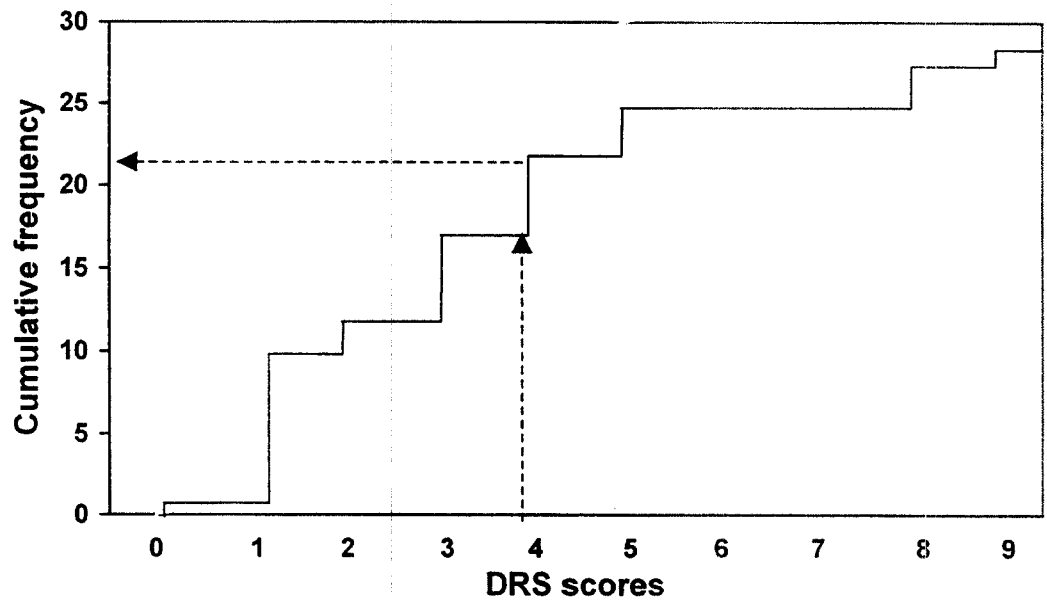


Figure 2.12 Step chart for cumulative frequency of DRS scores from Table 1.2.

Choice of technique

The table below provides a *guide* to choosing the most appropriate chart or table for the display of sample data.

	Type of data			
	Nominal categorical	Ordered categorical	Discrete metric	Continuous metric
To provide: ↓				
Graphical summary of frequency or value	Bar chart or pie chart*	Bar chart, dotplot or box plot†	Bar chart, dotplot or box plot	Histogram or box plot‡
Graphical summary of *cumulative* frequency		Step chart		Ogive
Tabular summary	Frequency table	Frequency, grouped frequency or cumulative frequency tables		Grouped frequency tables

* Pie charts work best when there are only 2 (at most 3) categories.
† Boxplots need at least 20 sample values.
‡ May be able to use a dotpolot if the data has been recorded as if discrete.

Solutions to Coursebook Questions

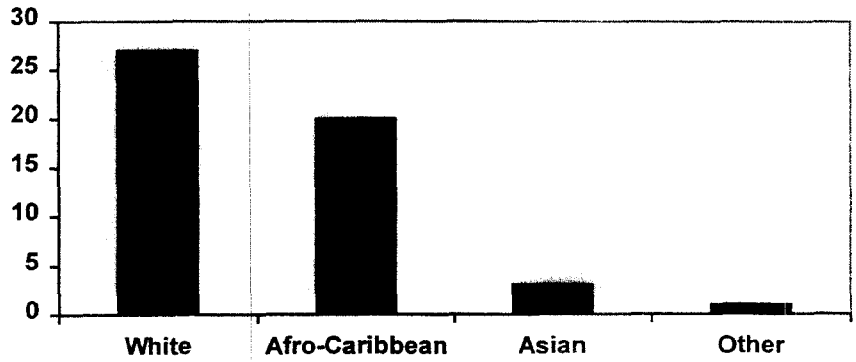
UNIT 2: Graphing Data

Q. 2.1 Positively skewed.

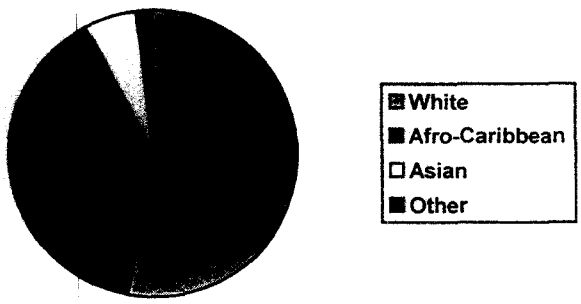
Q. 2.2 (a) Under 3 months (b) Over 12 months (no patients).

Q. 2.3 (a) "Want to be referred to specialist"; (b) "Want some tests done"; (c) "Want some test results" and "Want to know side effects", appear about the same.

Q. 2.4 Bar chart: bars drawn either in increasing or decreasing order of magnitude.



Pie chart: note, largest frequency is usually drawn to start at 0°.



Graphing data often provides insights which from the numbers alone may not be so obvious. Also a chart might make a more immediate impact than the corresponding table of data.

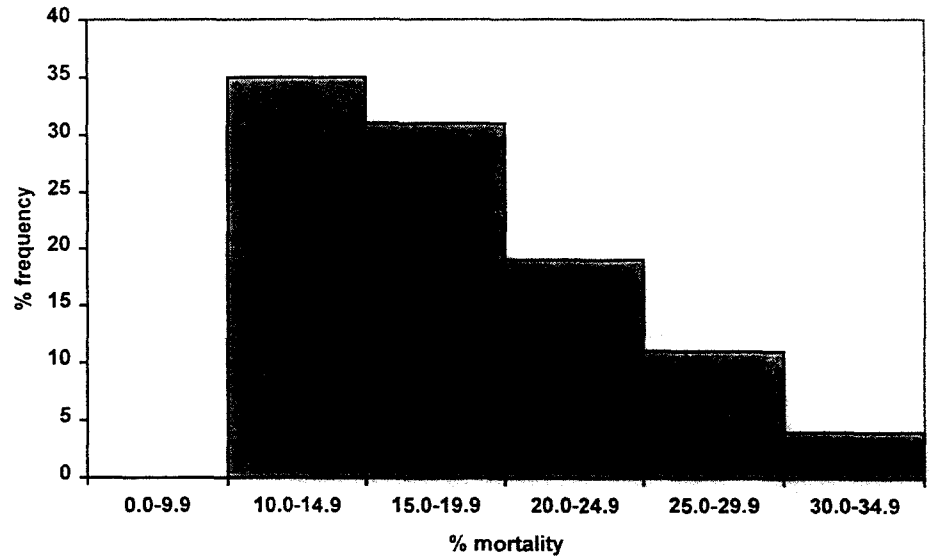
Q. 2.5 (a) Between 18 and 20; (b) positively skewed.

Q. 2.6 (a) Upper limit of oldest group is assumed to be 54 years. Since these women are attending a family planning clinic this seems reasonable.

Q. 2.7 Patient could be placed in either one of two categories or classes, 6-12 or 12-24 - not good! The categories must be designed so that every sample value can be placed in one and *only* one category or class. For example, <3 months; 3-6 months; 7-12 months; >12 months.

Q. 2.8 Not nearly as many patients in two uppermost groups as the incorrect histogram suggests, and distribution is much more positively skewed.

Q. 2.9 Comment: % mortality is positively skewed.



Q. 2.10 (a) 3 minutes (PCA) and 15 minutes (Traditional); (b) 2.25 minutes (PCA) and 10.08 minutes (Traditional); (c) 8 to 12½ minutes.

Q. 2.11 (a) about 75% and 81%; (b) both about 6mmol/l.