DESCRIPTIVE STATISTICS – DISPLAYING DATA

Several displays are possible to represent a set of data:

- frequency table: for categorical and numerical data
- bar graph: for categorical data
- Pareto chart
- two-way table
- divided bar graph
- dot plot: for discrete numerical or categorical data
- histogram: for continuous numerical data
- stem-and-leaf plot: for discrete numerical data

Frequency tables

A **frequency table** is used to summarise data. It can show frequency values for individual data values or grouped data. When grouped, the data may represent discrete data or continuous data.

Individual data

Number of TVs	Frequency
0	1
1	4
2	9
3	7
4	3

Grouped discrete data

Number of CDs	Frequency
0–9	3
10–19	5
20–29	6
30–39	2
40–49	1

Grouped continuous data

Distance travelled (km)	Frequency
0–<1	3
1–<2	5
2–<3	6
3–<4	2
4–<5	1

Bar graphs

A **bar graph** is used to display categorical (nominal) information. Values are simply read from the graph, using the frequency scale. Bar graphs can have vertical or horizontal bars, which are spaced out evenly and a gap is left between each of the categories. When vertical bars are used, the graph is sometimes called a column graph.

Example 1

Draw a bar graph to represent the following data collected as a result of a survey asking for the respondents' favourite code of football.

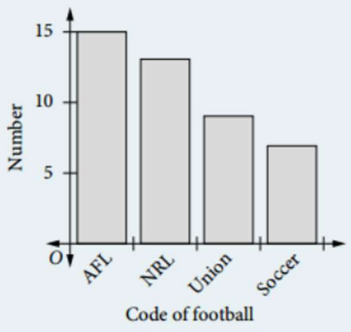
AFL	AFL	NRL	AFL	AFL	Soccer	AFL	NRL	NRL	AFL	Soccer
Soccer	NRL	Union	AFL	Union	NRL	Union	AFL	NRL	NRL	Union
Union	AFL	AFL	NRL	NRL	Union	AFL	Soccer	Union	NRL	NRL
AFL	Soccer	Soccer	AFL	Union	NRL	AFL	Union	NRL	AFL	Soccer

Solution

Construct a frequency table for the following data.

Football code	Frequency
AFL	15
NRL	13
Union	9
Soccer	7

Draw and label the bar graph, remembering to leave a gap between each of the categories.



Two-way tables

Two-way tables allow more information to be obtained from the table. As well as the information in each cell, you can also read information from each row or column.

Example 2

The foreign language studied by the students in a class are listed in the following table.

	Japanese	French	German	Total
Girls	6	5	4	15
Boys	2	7	6	15
Total	8	12	10	30

(a) How many students are in the class?

(b) How many girls are studying Japanese or German?

(c) How many students are studying French?

(d) What proportion of the students are not studying German?

Solution

(a) 30 students (the total number of girls and boys)

(b) Studying Japanese or German = $6 + 4 = 10$

(c) Studying French = $5 + 7 = 12$

(d) Studying German = 10

Not studying German = $30 - 10 = 20$

Proportion not studying German = $\frac{20}{30} = \frac{2}{3}$

Pareto charts

A Pareto chart is a type of chart that contains both bars and a line graph, where individual values are represented in descending order by the bars and the cumulative total is represented by the line graph. It is named after Vilfredo Pareto, an influential Italian sociologist and economist (although he originally trained as an engineer), who is also known for his work in politics and philosophy.

The left axis of the Pareto chart gives the frequency, cost or another unit of measure for the information shown in the bar graph, whilst the right axis shows the cumulative percentage total. By having the bars in descending order, the most frequent items are shown first.

The purpose of a Pareto chart is to distinguish the 'vital few' from the 'trivial many'. For example, if you are trying to analyse various different problems to decide which problems are the most important, then a Pareto chart can helpfully show which problems are having the greatest impact. The cumulative line will show how much of the total will be solved by fixing the few most frequent problems shown in the bar graph.

Example 3

An online seller of clothing summarised the complaints received in a month in the following table.

Type of complaint	Number of complaints
Problems completing the order online	50
Difficulty accessing website	30
Cancelled order	9
Wrong article sent	5
Overcharging for delivery	4
Late delivery	2
Total	**100**

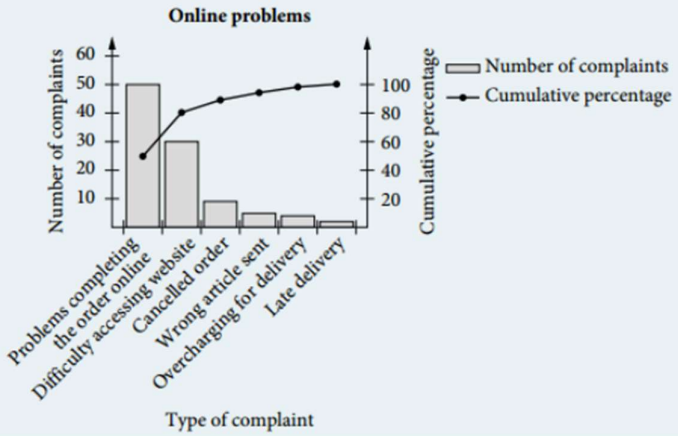
(a) Copy this table and add percentage and cumulative percentage columns.
(b) Construct a Pareto chart for this information.
(c) Which problems account for 80% of the complaints?

Solution

(a)

Complaint	Number of complaints	Percentage	Cumulative percentage
Problems completing the order online	50	50	50
Difficulty accessing website	30	30	80
Cancelled order	9	9	89
Wrong article sent	5	5	94
Overcharging for delivery	4	4	98
Late delivery	2	2	100
	100	**100**	

(b)



(c) 80% of the complaints occur in the categories 'Problems completing the order online' and 'Difficulty accessing the website'. These are the two areas that you would concentrate on fixing. If drawing a Pareto chart using spreadsheet software, it will help set the gridlines from the 'Cumulative percentage' axis as this will then clearly show the 80% line.

Example 4

A suggestion box is left at a train station for 5 days. The complaints received are summarised in the following table.

Type of complaint	Number of complaints
Crowded platform	1100
Lateness	1655
Crowded train	3290
Lack of toilets on platform	55
Dirty carriages	100
Total	**6200**

(a) Arrange the number of complaints in descending order and add percentage and cumulative percentage columns.

(b) Construct a Pareto chart for this information.

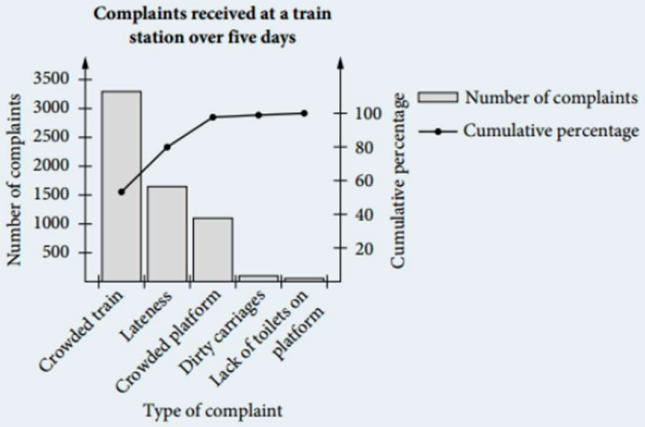
(c) What should the train company do to make the customers happier?

Solution

(a)

Type of complaint	Frequency	Percentage	Cumulative percentage
Crowded train	3290	53.1	53.1
Lateness	1655	26.7	79.8
Crowded platform	1100	17.7	97.5
Dirty carriages	100	1.6	99.1
Lack of toilets on platform	55	0.9	100
	6200	**100**	

(b)



(c) Most customers' problems would be solved by running more trains to reduce crowding and having them run on time.

Divided bar graphs

A **divided bar graph** or **segmented bar graph** is a rectangle divided into lengths according to the proportion of each group or category. Sometimes you can convert the values to percentages, rounded appropriately, to enable a simpler scale to be used.

Example 5

Draw a divided bar graph to represent the number of grams of various components in a particular brand of muesli bars.

Total per serving (g)	Protein (g)	Fat (g)	Carbohydrate (g)	Dietary fibre (g)	Other (g)
35	2.3	4.9	20.9	2.6	4.3

Solution

State the total grams per serving: There are 35 g in total.

Find the factors of the total (this will give convenient lengths for the rectangular whole):

$35 = 1 \times 35, 5 \times 7$

Select an appropriate length (approximately 10–20 cm where possible). Determine the length of each 'piece' by dividing the amount present by the number of grams per centimetre. Where necessary, state your answers to 1 decimal place:

Length: 7 cm

$$\frac{35}{7} = 5\,\text{g/cm (each centimetre represents 5 g)}$$

Protein: $\frac{2.3}{5} = 0.46\,\text{cm}$

Dietary fibre: $\frac{2.6}{5} = 0.52\,\text{cm}$

Fat: $\frac{4.9}{5} = 0.98\,\text{cm}$

Other: $\frac{4.3}{5} = 0.86\,\text{cm}$

Carbohydrate: $\frac{20.9}{5} = 4.18\,\text{cm}$

Draw a rectangle of a suitable length and divide it into pieces according to the category values. Include a key.

Dot plots

A **dot plot** can be used with either discrete numerical or categorical data. The categories or values are written along the horizontal axis, and then dots are arranged vertically to represent the number of each category. In this chapter you will only use dot plots for numerical data because categorical data can be better represented using a bar graph.

Example 6

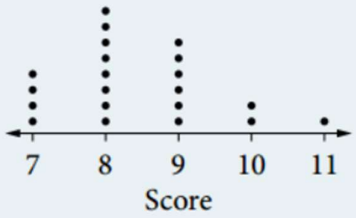
Draw a dot plot to represent the following discrete data set.

Score	Frequency
7	4
8	8
9	6
10	2
11	1
Total	**21**

Solution

Draw a horizontal line marked with the values from the 'Score' column:

For each score add the number of dots represented by the 'Frequency' value.



A dot plot is useful for identifying symmetry in a data distribution and for identifying extreme values that don't seem to fit with the other values. These values are called **outliers**.

A dot plot should not be used if there are too many values. It would be best to limit this type of display to no more than 10 columns. A dot plot can be used when the data is in its raw state. This means the data has not already been collated into a frequency table. In cases like this you should go through the data set once, marking dots in the appropriate places as you go.

A dot plot is also not very useful if there are high frequencies involved. As each value needs to be represented by its own dot you don't want to be drawing a dot plot where the individual frequencies are greater than 10.

Histograms

A **histogram** is used to display continuous numerical data. A histogram looks similar to a bar graph, but there are no gaps between the bars. Also, the first column is usually placed one half-column width from the vertical axis.

Example 7

Draw a histogram of the following data that represents the haemoglobin level for a sample of students.

Haemoglobin level	Frequency
9.0–<10	1
10.0–<11	2
11.0–<12	5
12.0–<13	2
13.0–<14	7
14.0–<15	5
15.0–<16	2
16.0–<17	1
Total	**25**

Solution

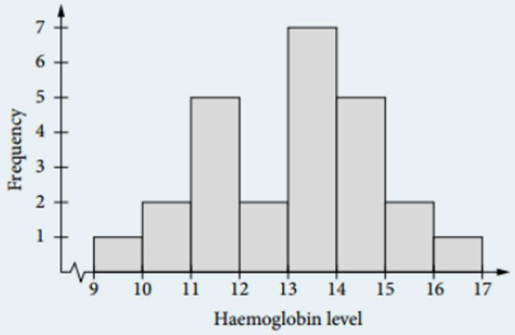
Find the minimum and maximum values to determine the range to be marked on the axes:

The minimum is greater than or equal to 9.

The maximum is less than 17.

The frequency is from 1 to 7.

Draw a set of axes marking 'Haemoglobin level' on the horizontal axis and 'Frequency' on the vertical axis. Draw a break at the start of the axis to indicate that part of the horizontal axis has been left out and complete the diagram.



Stem-and-leaf plots

A **stem-and-leaf plot** (or **stem plot**) is a useful way to display grouped discrete data. It does not lose any of the information, unlike for example a histogram (which does not show individual values).

A stem-and-leaf plot consists of two parts: a stem and a leaf. Typically, the leaf contains the last digit of the number and the stem contains all of the other digits. For example, the leaf may represent the 'ones'—single-digit numbers 1 to 9—and the stem will represent the rest of the number (tens, hundreds, thousands etc.). Use spaces, not commas, to separate each entry.

The values in the leaf section should be ordered. There should also be a key that explains the sort of values contained in the plot. The key will use a vertical line to show how the numbers are divided into stem and leaf.

The stem is usually based on a class interval of 10 but is also often based on 5. If based on 5, you can use subscript letters L (meaning *low*) and H (meaning *high*) to indicate where values belong. So, for a stem of 6, leaf values of 0, 1, 2, 3 and 4 would be placed alongside 6_L and leaf values of 5, 6, 7, 8 and 9 would be placed alongside 6_H, as in the following example that shows the data set 61, 61, 62, 64, 66, 68, 69, 69:

Stem	Leaf
6_L	1 1 2 4
6_H	6 8 9 9

Key: 6|4 = 64

A stem-and-leaf plot can also be used to record decimal values. For example, the following shows the times recorded for competitors in the 100 m run at an athletics carnival.

Stem	Leaf
10	9 9
11	0 1 4 5 6 8 9
12	0 1 1 1 2 4 5 6 9
13	0 0 1 1 2 2 2 2 3 4 6 7

Key: 10|9 = 10.9

Example 8

A class obtained the following marks, out of 100, on a mathematics test.

42, 55, 46, 78, 83, 85, 99, 78, 83, 66, 50, 49, 72, 84, 92, 81, 60, 56, 57, 82, 78, 57, 93, 58

Show the data in a stem-and-leaf plot with a class interval of 10.

Solution

Decide which values need to appear in the stem column. List them down the page in ascending order. Write the key you are using near the plot.

Work through the data in the order given, placing the leaf value in the appropriate row.

Stem	Leaf
4	2 6 9
5	5 0 6 7 7 8
6	6 0
7	8 8 2 8
8	3 5 3 4 1 2
9	9 2 3

Key: 5|0 = 50

Redraw the plot with the leaf values in order.

Stem	Leaf
4	2 6 9
5	0 5 6 7 7 8
6	0 6
7	2 8 8 8
8	1 2 3 3 4 5
9	2 3 9

Key: 5|0 = 50