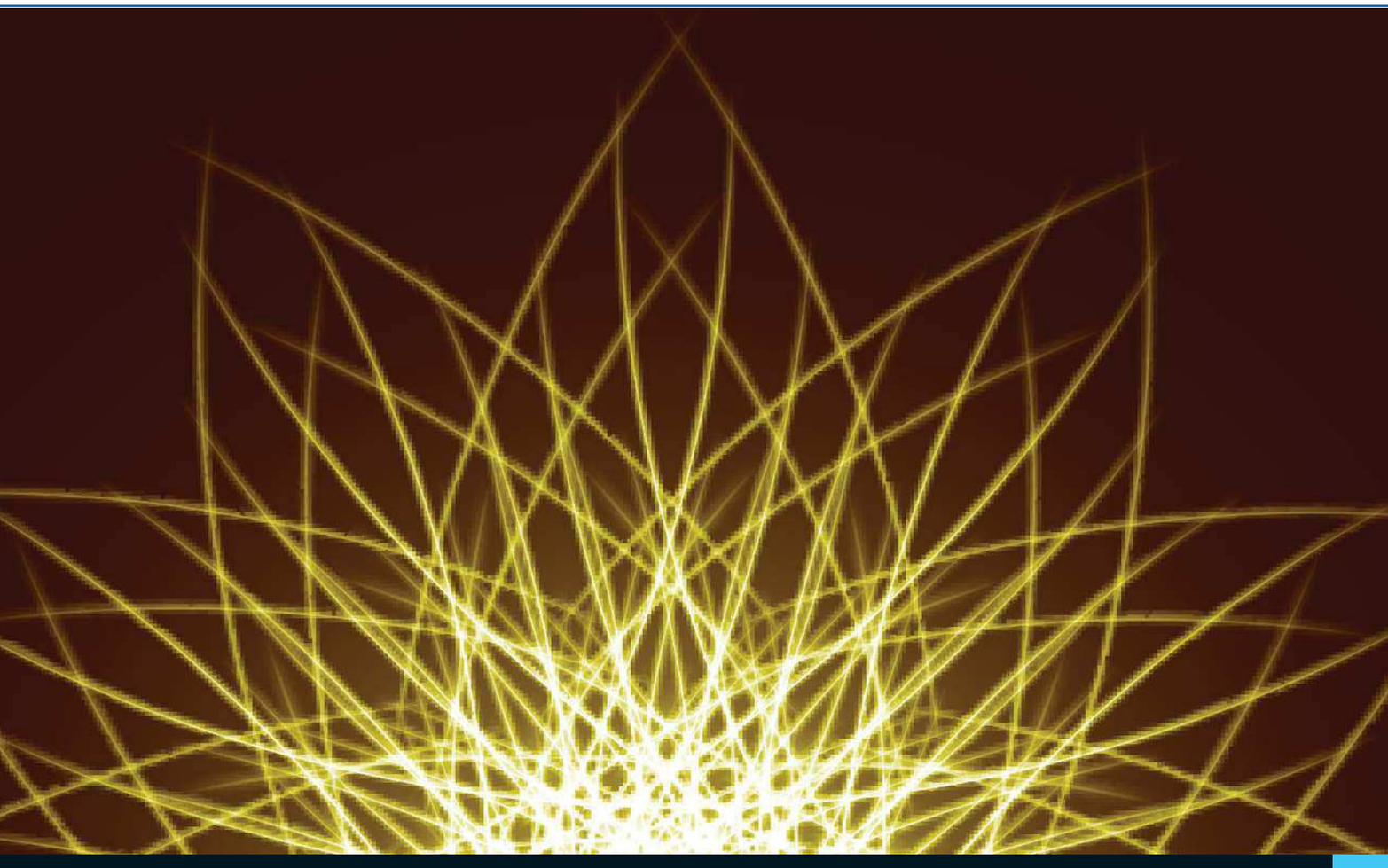

Chapter 1
Representation of data

This chapter looks at ways of displaying numerical data using diagrams. When you have completed it, you should

- know the difference between quantitative and qualitative data
- be able to make comparisons between sets of data by using diagrams
- be able to construct a stem-and-leaf diagram from raw data
- be able to draw a histogram from a grouped frequency table, and know that the area of each block is proportional to the frequency in that class
- be able to construct a cumulative frequency diagram from a frequency distribution table.

1.1 Introduction

The collection, organisation and analysis of numerical information are all part of the subject called **statistics**. Pieces of numerical and other information are called **data**. A more helpful definition of 'data' is 'a series of facts from which conclusions may be drawn'.

In order to collect data you need to observe or to measure some property. This property is called a **variable**. The data which follow were taken from the internet, which has many sites containing data sources. Electronic (Excel and CSV) versions of this data set are available on the Cambridge University Press website at http://www.cambridge.org/9781316600382.

In this example a variety of measurements was taken on packets of breakfast cereal in the United States. Each column of Table 1.1 represents a variable. So, for example, 'type', 'sodium' and 'shelf' are all variables. (The amounts for variables 3, 4, 5 and 6 are per serving.)

Datafile name Cereals

Description Data which refer to various brands of breakfast cereal in a particular store.

A value of –1 for nutrients indicates a missing observation.

Number of cases 77

Variable names

1. name: name of cereal
2. type: cold (C) or hot (H)
3. fat: grams of fat
4. sodium: milligrams of sodium
5. carbo: grams of complex carbohydrates
6. sugar: grams of sugars
7. shelf: display shelf (1, 2 or 3, counting from the floor)
8. mass: mass in grams of one serving
9. rating: a measure of the nutritional value of the cereal

Table 1.1 Datafile 'Cereals'

name	type	fat	sodium	carbo	sugar	shelf	mass	rating
100%_Bran	C	1	130	5	6	3	30	68
100%_Natural_Bran	C	5	15	8	8	3	30	34
All-Bran	C	1	260	7	5	3	30	59
All-Bran_with_Extra_Fiber	C	0	140	8	0	3	30	94
Almond_Delight	C	2	200	14	8	3	30	34
Apple_Cinnamon_Cheerios	C	2	180	10.5	10	1	30	30
Apple_Jacks	C	0	125	11	14	2	30	33
Basic_4	C	2	210	18	8	3	40	37
Bran_Chex	C	1	200	15	6	1	30	49
Bran_Flakes	C	0	210	13	5	3	30	53

name	type	fat	sodium	carbo	sugar	shelf	mass	rating
Cap'n'Crunch	C	2	220	12	12	2	30	18
Cheerios	C	2	290	17	1	1	30	51
Cinnamon_Toast_Crunch	C	3	210	13	9	2	30	20
Clusters	C	2	140	13	7	3	30	40
Cocoa_Puffs	C	1	180	12	13	2	30	23
Corn_Chex	C	0	280	22	3	1	30	41
Corn_Flakes	C	0	290	21	2	1	30	46
Corn_Pops	C	0	90	13	12	2	30	36
Count_Chocula	C	1	180	12	13	2	30	22
Cracklin'_Oat_Bran	C	3	140	10	7	3	30	40
Cream_of_Wheat_(Quick)	H	0	80	21	0	2	30	65
Crispix	C	0	220	21	3	3	30	47
Crispy_Wheat_&_Raisins	C	1	140	11	10	3	30	36
Double_Chex	C	0	190	18	5	3	30	44
Froot_Loops	C	1	125	11	13	2	30	32
Frosted_Flakes	C	0	200	14	11	1	30	31
Frosted_Mini-Wheats	C	0	0	14	7	2	30	58
Fruit_&_Fibre_Dates,_Walnuts,_and_Oats	C	2	160	12	10	3	40	41
Fruitful_Bran	C	0	240	14	12	3	40	41
Fruity_Pebbles	C	1	135	13	12	2	30	28
Golden_Crisp	C	0	45	11	15	1	30	35
Golden_Grahams	C	1	280	15	9	2	30	24
Grape_Nuts_Flakes	C	1	140	15	5	3	30	52
Grape-Nuts	C	0	170	17	3	3	30	53
Great_Grains_Pecan	C	3	75	13	4	3	30	46
Honey_Graham_Ohs	C	2	220	12	11	2	30	22
Honey_Nut_Cheerios	C	1	250	11.5	10	1	30	31
Honey-comb	C	0	180	14	11	1	30	29
Just_Right_Crunchy_Nuggets	C	1	170	17	6	3	30	37
Just_Right_Fruit_&_Nut	C	1	170	20	9	3	40	36
Kix	C	1	260	21	3	2	30	39
Life	C	2	150	12	6	2	30	45
Lucky_Charms	C	1	180	12	12	2	30	27
Maypo	H	1	0	16	3	2	30	55
Muesli_Raisins,_Dates,_&_Almonds	C	3	95	16	11	3	30	37
Muesli_Raisins,_Peaches,_&_Pecans	C	3	150	16	11	3	30	34
Mueslix_Crispy_Blend	C	2	150	17	13	3	45	30
Multi-Grain_Cheerios	C	1	220	15	6	1	30	40
Nut_&_Honey_Crunch	C	1	190	15	9	2	30	30

(*continued*)

Table 1.1 (*continued*)

name	type	fat	sodium	carbo	sugar	shelf	mass	rating
Nutri-Grain_Almond-Raisin	C	2	220	21	7	3	40	41
Nutri-Grain_Wheat	C	0	170	18	2	3	30	60
Oatmeal_Raisin_Crisp	C	2	170	13.5	10	3	40	30
Post_Nat._Raisin_Bran	C	1	200	11	14	3	40	38
Product_19	C	0	320	20	3	3	30	42
Puffed_Rice	C	0	0	13	0	3	15	61
Puffed_Wheat	C	0	0	10	0	3	15	63
Quaker_Oat_Squares	C	1	135	14	6	3	30	50
Quaker_Oatmeal	H	2	0	–1	–1	1	30	51
Raisin_Bran	C	1	210	14	12	2	40	39
Raisin_Nut_Bran	C	2	140	10.5	8	3	30	40
Raisin_Squares	C	0	0	15	6	3	30	55
Rice_Chex	C	0	240	23	2	1	30	42
Rice_Krispies	C	0	290	22	3	1	30	41
Shredded_Wheat	C	0	0	16	0	1	25	68
Shredded_Wheat'n'Bran	C	0	0	19	0	1	30	74
Shredded_Wheat_spoon_size	C	0	0	20	0	1	30	73
Smacks	C	1	70	9	15	2	30	31
Special_K	C	0	230	16	3	1	30	53
Strawberry_Fruit_Wheats	C	0	15	15	5	2	30	59
Total_Corn_Flakes	C	1	200	21	3	3	30	39
Total_Raisin_Bran	C	1	190	15	14	3	45	29
Total_Whole_Grain	C	1	200	16	3	3	30	47
Triples	C	1	250	21	3	3	30	39
Trix	C	1	140	13	12	2	30	28
Wheat_Chex	C	1	230	17	3	1	30	50
Wheaties	C	1	200	17	3	1	30	52
Wheaties_Honey_Gold	C	1	200	16	8	1	30	36

Computers are able to sort data very quickly and so make the process of producing stem and leaf diagrams for large data sets easier. Electronic (Excel and CSV) versions of the data sets shown in Tables 1.1 *and* 1.2 *are available on the Cambridge University Press website at http://www.cambridge.org/*9781316600382. *Students may wish to use them to check the stem and leaf diagrams in these pages or to make their own with some of the data not used.*

The variable 'type' has two different letter codes, H and C.

The variable 'sodium' takes values such as 130, 15, 260 and 140.

The variable 'shelf' takes values 1, 2 or 3.

You can see that there are different types of variable. The variable 'type' is non-numerical: such variables are usually called **qualitative**. The other two variables are called **quantitative**, because the values they take are numerical.

Quantitative data can be subdivided into two categories. For example, 'sodium', the mass of sodium in milligrams, which can take any value in a particular range, is called a **continuous** variable. 'Display shelf', on the other hand, is a **discrete** variable: it can only

take the integer values 1, 2 or 3, and there are clear steps between its possible values. It would not be sensible, for example, to refer to display shelf number 2.43.

In summary:

A variable is *qualitative* if it is not possible for it to take a numerical value.
A variable is *quantitative* if it can take a numerical value.
A quantitative variable which can take any value in a given range is *continuous*.
A quantitative variable which has clear steps between its possible values is *discrete*.

1.2 Stem-and-leaf diagrams

The datafile on cereals has one column which gives a rating of the cereals on a scale of 0–100. The ratings are given below.

68	34	59	94	34	30	33	37	49	53
18	51	20	40	23	41	46	36	22	40
65	47	36	44	32	31	58	41	41	28
35	24	52	53	46	22	31	29	37	36
39	45	27	55	37	34	30	40	30	41
60	30	38	42	61	63	50	51	39	40
55	42	41	68	74	73	31	53	59	39
29	47	39	28	50	52	36			

These values are what statisticians call **raw data**. Raw data are the values collected in a survey or experiment before they are categorised or arranged in any way. Usually raw data appear in the form of a list. It is very difficult to draw any conclusions from these raw data just by looking at the numbers. One way of arranging the values that gives some information about the patterns within the data is a **stem-and-leaf diagram**.

In this case the 'stems' are the tens digits and the 'leaves' are the units digits. You write the stems to the left of a vertical line and the leaves to the right of the line. So, for example, you would write the first value, 68, as 6|8. The leaves belonging to one stem are then written in the same row.

The stem-and-leaf diagram for the ratings data is shown in Fig. 1.2. The **key** shows what the stems and leaves mean.

0		(0)
1	8	(1)
2	0 3 2 8 4 2 9 7 9 8	(10)
3	4 4 0 3 7 6 6 2 1 5 1 7 6 9 7 4 0 0 0 8 9 1 9 9 6	(25)
4	9 0 1 6 0 7 4 1 1 6 5 0 1 2 0 2 1 7	(18)
5	9 3 1 8 2 3 5 0 1 5 3 9 0 2	(14)
6	8 5 0 1 3 8	(6)
7	4 3	(2)
8		(0)
9	4	(1)

Key: 6|8 means 68

Fig. 1.2 Unordered stem-and-leaf diagram of cereal ratings

The numbers in the brackets tell you how many leaves belong to each stem. The digits in each stem form a horizontal 'block', similar to a bar on a bar chart, which gives a visual impression of the distribution. In fact, if you rotate a stem-and-leaf diagram anticlockwise through 90°, it looks like a bar chart. It is also common to rewrite the leaves in numerical order; the stem-and-leaf diagram formed in this way is called an **ordered stem-and-leaf diagram**. The ordered stem-and-leaf diagram for the cereal ratings is shown in Fig. 1.3.

0		(0)
1	8	(1)
2	0 2 2 3 4 7 8 8 9 9	(10)
3	0 0 0 0 1 1 1 2 3 4 4 4 5 6 6 6 6 7 7 7 8 9 9 9 9	(25)
4	0 0 0 0 1 1 1 1 1 2 2 4 5 6 6 7 7 9	(18)
5	0 0 1 1 2 2 3 3 3 5 5 8 9 9	(14)
6	0 1 3 5 8 8	(6)
7	3 4	(2)
8		(0)
9	4	(1)

Key: 6|8 means 68

Fig. 1.3 Ordered stem-and-leaf diagram of cereal ratings

So far the stem-and-leaf diagrams discussed have consisted of data values that are integers between 0 and 100. With suitable adjustments, you can use stem-and-leaf diagrams for other data values.

For example, the data 6.2, 3.1, 4.8, 9.1, 8.3, 6.2, 1.4, 9.6, 0.3, 0.3, 8.4, 6.1, 8.2, 4.3 could be illustrated in the stem-and-leaf diagram in Fig. 1.4.

0	3 3	(2)
1	4	(1)
2		(0)
3	1	(1)
4	8 3	(2)
5		(0)
6	2 2 1	(3)
7		(0)
8	3 4 2	(3)
9	1 6	(2)

Key: 0|3 means 0.3

Fig. 1.4 Stem-and-leaf diagram

Table 1.5 is a datafile about brain sizes which will be used in several examples. Electronic (Excel and CSV) versions of this data set are available on the Cambridge University Press website at http://www.cambridge.org/9781316600382.

Datafile name Brain size

(Data from *Intelligence*, Vol. 15, Willerman et al., 'In vivo brain size ...', 1991)

Description A team of researchers used a sample of 40 students at a university. The subjects took four subtests from the 'Wechsler (1981) Adult Intelligence Scale – Revised' test. Magnetic resonance imaging (MRI) was then used to measure the brain sizes of the subjects. The subjects' genders, heights and body masses are also

included. The researchers withheld the masses of two subjects and the height of one subject for reasons of confidentiality.

Number of cases 40

Variable names

1. gender: male or female
2. FSIQ: full scale IQ scores based on the four Wechsler (1981) subtests
3. VIQ: verbal IQ scores based on the four Wechsler (1981) subtests
4. PIQ: performance IQ scores based on the four Wechsler (1981) subtests
5. mass: body mass in kg
6. height: height in cm
7. MRI_Count: total pixel count from 18 MRI brain scans

Table 1.5 Datafile 'Brain size'

gender	FSIQ	VIQ	PIQ	mass	height	MRI_Count
Female	133	132	124	54	164	816 932
Male	140	150	124	—	184	1 001 121
Male	139	123	150	65	186	1 038 437
Male	133	129	128	78	175	965 353
Female	137	132	134	67	165	951 545
Female	99	90	110	66	175	928 799
Female	138	136	131	63	164	991 305
Female	92	90	98	79	168	854 258
Male	89	93	84	61	168	904 858
Male	133	114	147	78	175	955 466
Female	132	129	124	54	164	833 868
Male	141	150	128	69	178	1 079 549
Male	135	129	124	70	175	924 059
Female	140	120	147	70	179	856 472
Female	96	100	90	66	168	878 897
Female	83	71	96	61	173	865 363
Female	132	132	120	58	174	852 244
Male	100	96	102	81	187	945 088
Female	101	112	84	62	168	808 020
Male	80	77	86	82	178	889 083
Male	83	83	86	—	—	892 420
Male	97	107	84	84	194	905 940
Female	135	129	134	55	157	790 619
Male	139	145	128	60	173	955 003
Female	91	86	102	52	160	831 772
Male	141	145	131	78	183	935 494
Female	85	90	84	64	173	798 612

(*continued*)

Table 1.5 (*continued*)

gender	FSIQ	VIQ	PIQ	mass	height	MRI_Count
Male	103	96	110	85	196	1 062 462
Female	77	83	72	48	160	793 549
Female	130	126	124	72	169	866 662
Female	133	126	132	58	159	857 782
Male	144	145	137	87	170	949 589
Male	103	96	110	87	192	997 925
Male	90	96	86	82	175	879 987
Female	83	90	81	65	169	834 344
Female	133	129	128	69	169	948 066
Male	140	150	124	65	179	949 395
Female	88	86	94	63	164	893 983
Male	81	90	74	67	188	930 016
Male	89	91	89	81	192	935 863

The stems of a stem-and-leaf diagram may consist of more than one digit. So, for example, consider the following data, which are the heights of 39 people in cm (correct to the nearest cm), taken from the datafile 'Brain size'.

164	184	186	175	165	175	164	168	168	175	164	178	175
179	168	173	174	187	168	178	194	157	173	160	183	173
196	160	169	159	170	192	175	169	169	179	164	188	192

You can represent these data with the stem-and-leaf diagram shown in Fig. 1.6, which uses stems from 15 to 19.

15	7 9	(2)
16	0 0 4 4 4 4 5 8 8 8 8 9 9 9	(14)
17	0 3 3 3 4 5 5 5 5 5 8 8 9 9	(14)
18	3 4 6 7 8	(5)
19	2 2 4 6	(4)

Key: 17|3 means 173 cm

Fig. 1.6 Stem-and-leaf diagram of heights of a sample of 39 people

Sometimes, statisticians want to compare two sets of data to understand their similarities and their differences. One way to represent two sets of data which are measuring the same property for two different groups is to use a **back-to-back stem-and-leaf diagram**. For example, if you want to compare the heights of the males and females in the 'Brain size' data, you could create a back-to-back stem-and-leaf diagram as shown in Fig. 1.7. This is similar to a normal stem-and-leaf diagram, except that:

a the stems are written in the middle, and the leaves are to the left and right;

b the leaves on the right are written in increasing order, as with a normal ordered stem-and-leaf diagram, while those on the left are written in decreasing order; in other words, the smallest leaves are always placed closest to the stem;

c the left and right sides need titles, so that it is clear which group each refers to;

d the key must explain what the leaves on both sides mean.

	Males		Females	
(0)		15	7 9	(2)
(1)	8	16	0 0 4 4 4 4 5 8 8 8 9 9 9	(13)
(9)	9 8 8 5 5 5 5 3 0	17	3 3 4 5 9	(5)
(5)	8 7 6 4 3	18		(0)
(4)	6 4 2 2	19		(0)

Key: 3 | 17 | 5 means 173 cm for the Males and 175 cm for the Females

Fig. 1.7 Back-to-back stem-and-leaf diagram of heights of a sample of 39 people

Note that a back-to-back stem-and-leaf diagram only makes sense when comparing the same sort of data about two different groups.

Exercise 1A

In this exercise, if you are asked to construct a stem-and-leaf diagram you should give an ordered one.

1 The following stem-and-leaf diagram illustrates the lengths, in cm, of a sample of 15 leaves fallen from a tree. The values are given correct to 1 decimal place.

4	3	(1)
5	4 0 7 3 9	(5)
6	3 1 2 4	(4)
7	6 1 6	(3)
8		(0)
9	3 2	(2)

Key: 7|6 means 7.6 cm

a Write the data in full, and in increasing order of size.

b State whether the variable is (i) qualitative or quantitative, (ii) discrete or continuous.

2 Construct stem-and-leaf diagrams for the following data sets.

a The speeds, in kilometres per hour, of 20 cars, measured on a city street.

41 15 4 27 21 32 43 37 18 25 29 34 28 30 25 52 12 36 6 25

b The times taken, in hours (to the nearest tenth), to carry out repairs to 17 pieces of machinery.

0.9 1.0 2.1 4.2 0.7 1.1 0.9 1.8 0.9 1.2 2.3 1.6 2.1 0.3 0.8 2.7 0.4

3 Construct a stem-and-leaf diagram for the following ages (in completed years) of famous people with birthdays on June 14 and June 15, as reported in a national newspaper.

75 48 63 79 57 74 50 34 62 67 60 58 30 81 51 58 91 71 67 56 74
50 99 36 54 59 54 69 68 74 93 86 77 70 52 64 48 53 68 76 75 56

4 The tensile strength of 60 samples of rubber was measured and the results, in suitable units, were as follows.

174 160 141 153 161 159 163 186 179 167 154 145 156 159 171
156 142 169 160 171 188 151 162 164 172 181 152 178 151 177
180 186 168 169 171 168 157 166 181 171 183 176 155 161 182
160 182 173 189 181 175 165 177 184 161 170 167 180 137 143

Construct a stem-and-leaf diagram using two rows for each stem so that, for example, with a stem of 15 the first leaf may have digits 0 to 4 and the second leaf may have digits 5 to 9.

5 A selection of 25 of A. A. Michelson's measurements of the speed of light, carried out in 1882, is given below. The figures are in thousands of kilometres per second and are given correct to 5 significant figures.

299.84	299.96	299.87	300.00	299.93	299.65	299.88	299.98	299.74
299.94	299.81	299.76	300.07	299.79	299.93	299.80	299.75	299.91
299.72	299.90	299.83	299.62	299.88	299.97	299.85		

Construct a suitable stem-and-leaf diagram for the data.

6 The contents of 30 medium-size packets of soap powder were weighed and the results, in kilograms correct to 4 significant figures, were as follows.

1.347	1.351	1.344	1.362	1.338	1.341	1.342	1.356	1.339	1.351
1.354	1.336	1.345	1.350	1.353	1.347	1.342	1.353	1.329	1.346
1.332	1.348	1.342	1.353	1.341	1.322	1.354	1.347	1.349	1.370

a Construct a stem-and-leaf diagram for the data.

b Why would there be no point in drawing a stem-and-leaf diagram for the data rounded to 3 significant figures?

7 Two teenagers kept a record of how many texts they had sent each day for a month. The results were as follows.

Teenager A:

97	79	73	87	60	85	82	86	72	96
68	79	103	88	62	72	78	77	76	81
79	91	90	87	80	71	93	77	69	70

Teenager B:

90	85	92	82	97	79	97	72	92	85
78	78	80	100	75	72	79	107	112	97
113	97	94	96	92	90	105	68	87	97

Construct a back-to-back stem-and-leaf diagram for the data.

1.3 Histograms

Sometimes, for example if the data set is large, a stem-and-leaf diagram is not the best method of displaying data, and you need to use other methods. For large sets of data you may wish to divide the data into groups, called classes.

In the data on cereals, the amounts of sodium may be grouped into classes as in Table 1.8.