

# Exploring Data

36.2



## Introduction

Techniques for exploring data to enable valid conclusions to be drawn are described in this Section. The diagrammatic methods of stem-and-leaf and box-and-whisker are given prominence. You will also learn how to summarize data using sets of statistics which have meaning in cases where a data set is not symmetrical. You should note that statistics such as the mean and variance are of limited use in such situations. Finally, you will encounter outliers. These are values which lie outside the main body of the data set and can enable you to reach important conclusions about the behaviour of the data.



## Prerequisites

Before starting this Section you should ...

- ① understand the ideas of sets and subsets. (section 35.1)



## Learning Outcomes

After completing this Section you should be able to ...

- ✓ understand the principles of EDA (Exploratory Data Analysis)
- ✓ be able to construct stem-and-leaf diagrams and box-and-whisker plots
- ✓ understand the significance of outliers, skewness, gaps and multiple peaks

# 1. Exploratory Data Analysis

## Introduction

The title ‘Exploratory Data Analysis’ (EDA) is usually taken to mean the activity by which data is explored and organized in order that information it contains is made clear. This branch of statistics usually deals with summary statistics which are resistant to departures from normality. The techniques used in EDA were first developed by the statistician John Tukey and for details of EDA which are beyond this open learning booklet, you are referred to the text *Exploratory Data Analysis*, by J.W. Tukey, Addison-Wesley, 1977. Tukey’s techniques have been used in innumerable papers and books since that date.

## The Basics of EDA

The basic principles followed in EDA are:

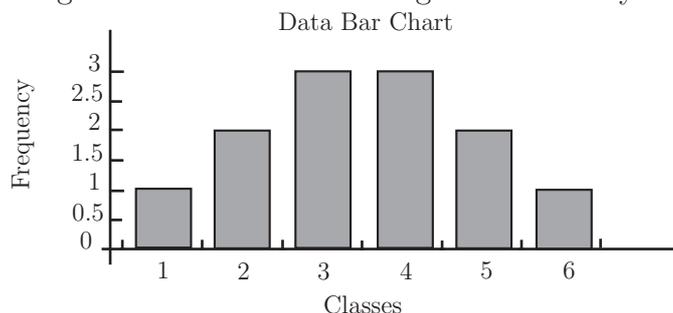
- To measure the location and spread of a distribution we use statistics which are **resistant** to departures from normality;
- To summarise shape location and spread we use several statistics rather than just two;
- Visual displays as well as numerical displays are used to summarise information obtained about shape, location and spread.

You can see these principles illustrated below.

Traditionally, the location and spread of a distribution are measured by calculating its mean and standard deviation. The problem with these statistics is that are sensitive to the influence of extreme values. For example, the data set

1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 6

has mean  $\mu = 3.5$  and standard deviation  $\sigma = 1.46$ . These values are quite acceptable since the distribution is symmetrical about its mean of 3.5. The symmetry is easily seen simply by inspecting the data although the bar chart below might make the symmetry more obvious.



The shape of the distribution may also be shown by the **stem-and-leaf** diagram below. Notice that the *stem* consists of the numbers 1 to 6 and the *leaves* are just the members of each class.

```
1 | 1
2 | 2 2
3 | 3 3 3
4 | 4 4 4
5 | 5 5
6 | 6
```

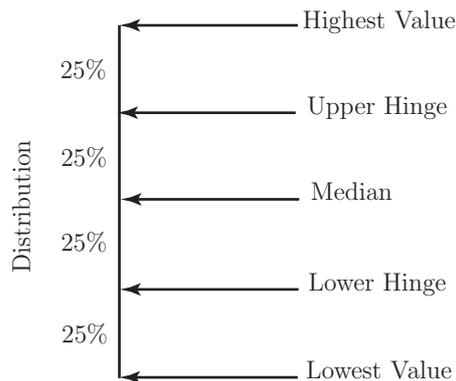
You will study the stem-and-leaf diagram in more detail later in this booklet.

The effects of changes in extreme values are easily illustrated by looking at what happens if we take the last number to be 60 instead of 6. This destroys the symmetry of the distribution and gives mean  $\mu = 8$  and standard deviation  $\sigma = 16.42$ .

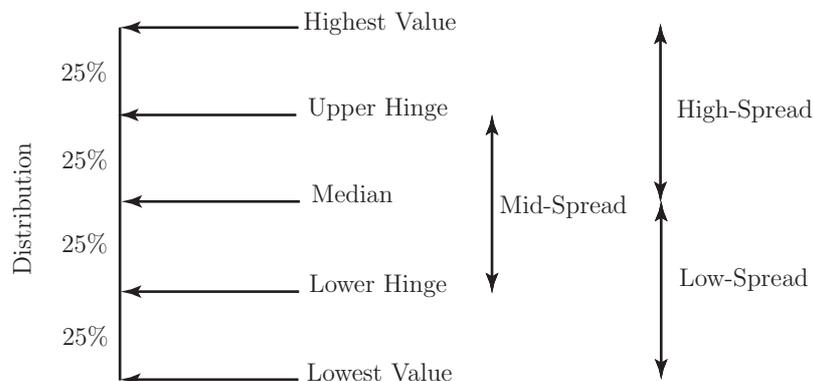
Clearly, these values do not describe the distribution very well at all, a mean which is higher than 92% of the members of the distribution can hardly be described as representative!

The simplest and most common examples of resistant statistics are those based on the idea of rank order - we simply order a distribution starting at the highest value and ending at the lowest value (or lowest to highest).

The five essential statistics based on rank order are illustrated in the diagram below:



Using these values other statistics which represent the shape or spread of the distribution may be defined. These statistics are known as the Mid-Spread, High-Spread and Low-Spread and their definition is indicated in the diagram below.



Elementary EDA recommends the use of a **five-number summary** consisting of:

1. the lowest value;
2. the lower hinge;
3. the median;
4. the upper hinge;
5. the highest value.

to summarize a distribution. You will find that the five-number summary, especially when used in conjunction with the three spreads shown in the diagram above gives an adequate representation of a non-symmetrical distribution.

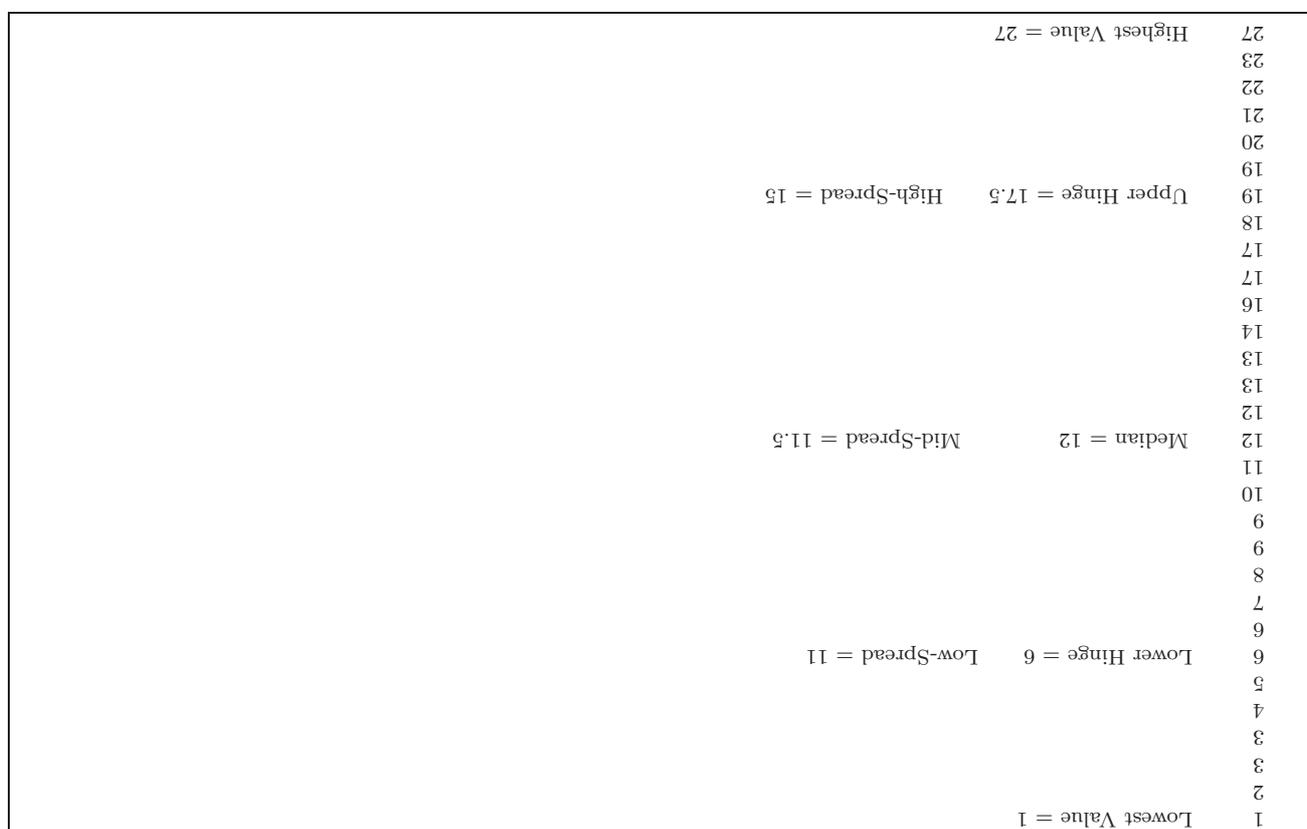
Notice that:

- the spreads shown in the diagram above are easily calculated once the five-number summary is known;
- the median and the hinges are unaffected by changes in extreme values.



Find the five number summary and the mid-spread, high-spread and low-spread for the distribution given below.

1 9 17 2 9 17 3 10 18 3 11 19 4 12 19  
5 12 20 6 13 21 6 13 22 7 14 23 8 16 27



## The Stem-and-Leaf Diagram

You have already seen a basic stem-and-leaf diagram and you know that it shows the shape of a distribution well. Here you will learn how to handle larger amounts of data to form stem-and-leaf diagrams. As you will see, one set of data can give rise to more than one stem-and-leaf diagram and highlight different aspects of the data. Look at the data set below:

11 9 6 27 17 2 19 12 8 17 3 10 23 6 18  
13 11 22 13 19 4 12 23 34 19 15 7 40 16 20

Using the numbers to the left of the stem to represent 10s and the numbers to the right to represent units we obtain the stem-and-leaf diagram shown below.



Notice that the skewed nature of the data stands out immediately. What also stands out are the following:

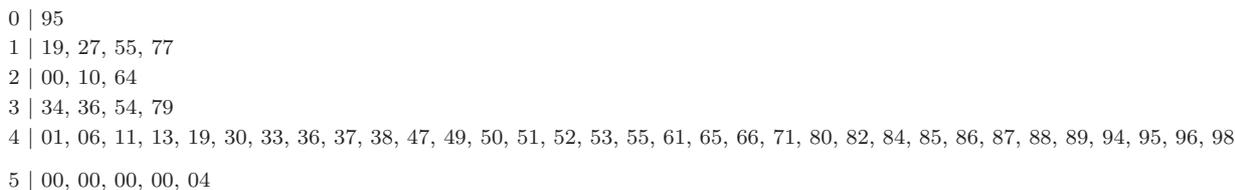
- the 10s class has the highest number of members;
- the modal (most frequently occurring) value is 19;
- the 30s and 40s tie for the least number of members (one each).

This is not new information, we could have written these fact down after properly inspecting the original raw data. The advantage of the stem-and-leaf diagram is that it enables these facts to be expressed in a clear and obvious way. As a further illustrative example, look at the data in the table below which we will use to draw two stem-and-leaf diagrams.

9.5	11.9	20.0	33.4	40.1	50.0	12.7	21.0	33.6	40.6
50.0	15.5	26.4	35.4	41.1	50.0	17.7	37.9	41.3	50.0
41.9	50.4	43.0	43.3	43.6	43.7	43.8	44.7	44.9	45.0
45.1	45.2	45.3	45.5	46.1	46.5	46.6	47.1	48.0	48.2
48.5	48.4	48.6	48.7	48.8	48.9	49.4	49.5	49.6	49.8

## Drawing a Stem-and-Leaf Diagram

We can start by looking at the data as it is displayed by a stem-and-leaf diagram. Here we will use two-digit leaves with the first digit representing units and the second digit representing tenths. The tens are represented by the numbers to the left of the stem.



Notice that all we have really done is rank the data from the lowest value to the highest value reading from top to bottom. This particular display has over half of its members crushed into one class - the 4-class. It may be informative to split the classes and look more closely at the data.

This can be done by:

1. rounding the raw data to two figures;
2. splitting each class according to the rule

second digit 0 - 4 ..... \*

second digit 5 - 9 ..... ●

The rounded raw data now appears as follows

10	12	20	33	40	50	13	21	34	41
50	16	26	35	41	50	18	38	41	50
42	50	43	43	44	44	44	45	45	45
45	45	45	46	46	47	47	47	48	48
49	48	49	49	49	49	49	50	50	50

The stem and leaf diagram now becomes

0*																				
0•																				
1*	0	2	3																	
1•	6	8																		
2*	0	1																		
2•	6																			
3*	3	4																		
3•	5	8																		
4*	0	1	1	1	2	3	3	4	4	4										
4•	5	5	5	5	5	5	6	6	7	7	7	8	8	8	9	9	9	9	9	9
5*	0	0	0	0	0	0	0	0	0											

Essentially, the classes have been split according to the usual rule for rounding decimals. This process can make certain information contained in the data a little more obvious than the previous stem and leaf diagram. For example:

- the values in the 3-class are evenly distributed between both halves of the class in the sense that each half has two members;
- the 4-class is split in the ratio 2:1 in favour of the upper half of the class;
- the values in the 5-class are all in the lower half of the class.

You should have realised that:

- this is not *new* information - the new display has merely highlighted certain aspects of the raw data;
- some of the conclusions may have been affected by the rounding process.

Looking at the original stem and leaf diagram of the Inter-party data it is easy to produce a five-number summary of the data. The summary is:

1. The lowest value, this is 9.50;
2. The lower hinge, this is 39 (to find the lower hinge average the 12th and 13th values);
3. The median, this is 45.05 (the average of the 25th and 26th values);
4. The upper hinge, this is 48.55 (to find the upper hinge average the 37th and 38th values);
5. The highest value, this is 50.4.

The corresponding spreads are:

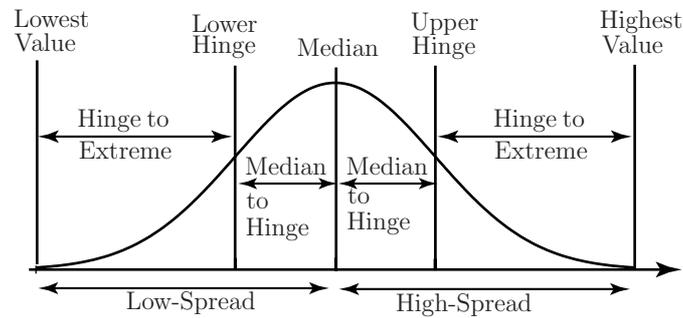
1. The low-spread, this is  $45.05 - 9.50 = 35.55$ ;
2. The mid-spread, this is  $48.55 - 39.00 = 9.55$ ;
3. The high-spread, this is  $50.40 - 45.05 = 5.35$ .

Notice that the spreads indicate a considerable deviation from normality. For an ideal normal distribution, we would expect:

- The distances between the median and hinges to be equal

- The high-spread and low-spread to be equal
- The distances between the hinges and the extremes to be equal

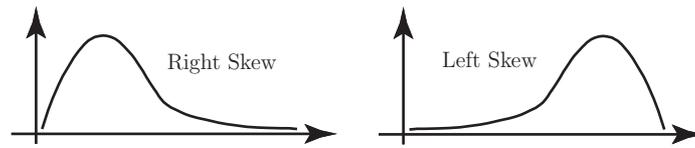
as shown in the following diagram.



Using the rounded data given above find its five number summary. Use your summary to check the data for normality and comment on any deviations from normality that you find.

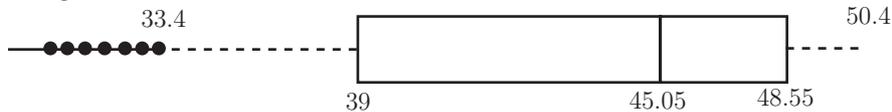


Remember that the term ‘skewness’ refers to the location of the ‘tail’ of a distribution.



## The Box-and-Whisker Diagram

In order to visually summarise a data set we can use a **box and whisker** plot as well as a stem-and-leaf diagram. A box-and-whisker diagram of the original (unrounded) Inter-Party Competition data is shown below and the procedure necessary for drawing a plot is discussed. You should note that there are several similar methods recommended by different authors for drawing box-and-whisker plots and so the methods recommended in statistical texts may vary a little from those given below.



The diagram is constructed as follows:

### 1. The Box

- (a) The left-hand vertical is placed at the lower hinge (39);
- (b) The right-hand vertical is placed at the upper hinge (48.65);
- (c) The vertical in the box is placed at the median (45.05).

### 2. The Whiskers

Notice that the mid-spread of the data (the difference between the hinges) is 9.65.

- (a) Find the greatest value which is within one mid-spread (9.65) of the upper hinge (48.65). Here  $48.65 + 9.65 = 58.3$  so the greatest value is 50.4.
- (b) Find the least value which is within one mid-spread (9.65) of the lower hinge (39). Here  $39 - 9.65 = 29.35$  so the least value is 33.4.

Connect the greatest and least values to the box by means of dashed lines.

### 3. Outlying Values

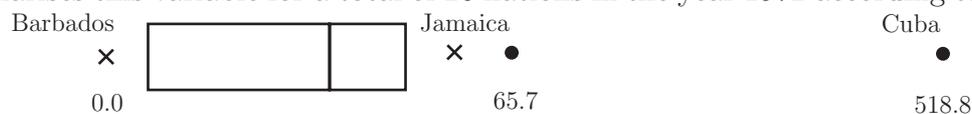
Mark as large dots any values which are **more** than 1.5 mid-spreads from the hinges. In this case 1.5 mid-spreads give a value of about 14.33 and so we mark dots which represent values which are higher than  $48.65 + 14.33 = 62.88$  and values which are lower than  $39 - 14.33 = 24.67$ . In this example there are no values greater than 62.88, but there are 7 values which are less than 24.67.

Notice that half of the data values lie in the box and that the tails show up well in the diagram. The diagram shows the left-skew (skewness refers to the tail) present in the data.

## 2. Outliers

Outliers are values which are well outside the range covered by the vast bulk of a data set - a precise definition is impossible although some simple criteria do exist which may be used to detect outliers and accept or reject outliers. The seven values shown as large dots above illustrate the concept of outliers. Outliers can be extremely important since they may be (for example) erroneous data or they may point the way to further investigations of a data set.

For example, one statistic used to measure the state of the industrial development of a nation is the number of miles of railway track built per square mile of land. The box-and-whisker plot below summarises this variable for a total of 26 nations in the year 1972 according to one author.



The figure for Cuba literally means that the whole island is covered by tracks which are placed about 3m apart! Clearly, there is an error in the data. In fact the 1972 Statistical Abstract of Latin America gives the figure for Cuba as 71.75 miles of railway per square mile of land. Note that the figure is still an outlier but is much more believable.



Place the items in the data set below in rank order and use your rank ordering to find the five number summary of the data.

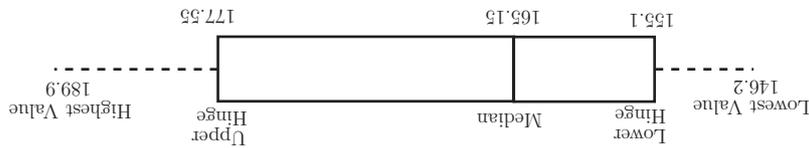
155.3 177.3 146.2 163.1 161.8 146.3 167.9 165.4 172.3 188.2  
 178.8 151.1 189.4 164.9 174.8 160.2 187.1 163.2 147.1 182.2  
 178.2 172.8 164.4 177.8 154.6 154.9 176.3 148.5 161.8 178.4

Construct a box-and-whisker diagram representing the data.

Does the box-and-whisker diagram tell you that the data set that you are working with is symmetrical? Record the reasons for your comments.

Data	
146.2	Lowest Value = 146.2
146.3	
147.1	
148.5	
151.1	
154.6	Lower Hinge = 155.1
154.9	
155.3	
160.2	
161.8	
161.8	
161.8	
163.1	
163.2	
164.4	
164.9	Median = 165.15
165.4	
167.9	
172.3	
172.8	
174.8	
176.3	
177.3	
177.8	Upper Hinge = 177.55
178.2	
178.4	
178.8	
182.2	
187.1	
188.2	
189.4	Highest Value = 189.4

The plot indicates that the distribution is not symmetrical, for example you would expect the median value to appear midway between the hinges for a symmetrical distribution.



## Criteria for Rejecting Outliers

As you already know, outliers may be taken to be observations which lie well outside the range of most of a sample. They are important for several reasons:

1. they can have misleading effect on statistics such as the mean and standard deviation;
2. their occurrence may be due to incorrect observation, measurement or recording. In this case it is often possible to correct the data;
3. their presence can induce a false skewness in a data set;
4. they may actually be members of a population not under consideration. For example, a study of urban families may involve recording the number of children in a family, say between 0 and 4 for the sake of discussion. An outlier might be caused by a rural family with, say, 10 children, living in temporary urban accommodation. This family is part of a different population.

Simple criteria exist which facilitate the detection of outliers. These criteria should be used with some caution and never automatically used simply to reject an outlier. You should always ask why such a value occurred in the first place and work to answer such a question sensibly before

considering rejection. Two criteria for the detection of outliers are given below. Criterion 1 may be applied to data sets that are known to be normal in shape. Criterion 2 uses the five-number summary discussed above and may be applied to any data sets.

### Criterion 1

Knowing that some 99.7% of a normal population lies within 3 standard deviations of the mean, we could treat any value further than say 3.3 standard deviations from the mean as an outlier. This choice essentially implies that a value has less than 1 in a 1000 chance of occurring naturally outside the range defined by 3.3 standard deviations from the mean. Using standardized scores with a potential outlier we can state the criterion

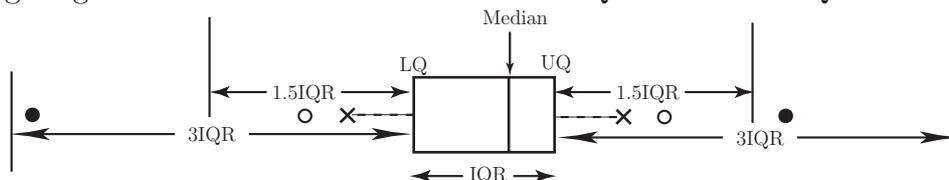
$$\text{Accept } x_0 \text{ if } \left| \frac{x_0 - \mu}{\sigma} \right| \leq 3.3 \quad \text{Investigate } x_0 \text{ if } \left| \frac{x_0 - \mu}{\sigma} \right| > 3.3$$

Note that  $\mu$  and  $\sigma$  are the mean and standard deviation of the whole data set (the 'population') under consideration.

### Criterion 2

Using a five-number summary of a data set one can easily set up a criterion which may be used to classify outliers as either 'moderate' or 'extreme'.

The following diagram illustrates the situation where IQR is the Inter-Quartile Range.



While all values classified as outliers should be investigated, this is particularly true of those classified as extreme outliers.



Manufacturing processes generally result in a certain amount of wasted material. For reasons of cost, companies need to keep such wastage to a minimum. The following data were gathered over a two week period by a manufacturing company whose production lines run seven days per week. The figures given represent the percentage wastage of the amount of material used in the manufacturing process.

Daily Losses (%)    6   8   10   12   12   13   14   14   18   18   19   20   22   26

Find the mean and standard deviation of the percentage losses of material over the two week period.

Assuming that the losses are roughly normally distributed, apply an appropriate criterion to decide whether any of the losses are smaller or larger than might be expected by chance.

We will treat any value further than 3.3 standard deviations from the mean as an outlier (criterion 1). Using standardized scores with  $x_0$  as the potential outlier we need to calculate the quantity

$$\left| \frac{x_0 - \bar{x}}{s} \right|$$

and then accept  $x_0$  as a member of the distribution if  $\left| \frac{x_0 - \bar{x}}{s} \right| \leq 3.3$ . Otherwise we reject  $x_0$  as an outlier. Calculation gives:

Data	$x - \bar{x}$	$(x - \bar{x})^2$	$\left  \frac{x - \bar{x}}{s} \right $
6.00	-9.14	83.59	1.69
8.00	-7.14	51.02	1.32
10.00	-5.14	26.45	0.95
12.00	-3.14	9.88	0.58
12.00	-3.14	9.88	0.58
14.00	-1.14	1.31	0.21
14.00	-1.14	1.31	0.21
18.00	2.86	8.16	0.53
18.00	2.86	8.16	0.53
19.00	3.86	14.88	0.71
20.00	4.86	23.59	0.90
22.00	6.86	47.02	1.27
26.00	10.86	117.88	2.01
$\bar{x} = 15.14$			
$s = 5.40$			

The calculation shows that all values of  $\left| \frac{x_0 - \bar{x}}{s} \right| \leq 3.3$  and so we conclude that the daily losses are within the range indicated by chance variation.

### 3. Skewness, Gaps and Multiple Peaks

When exploring a data set, four properties worth looking for are outliers, skewness, gaps and multiple peaks. Outliers have been dealt with in some detail above so the comments given below briefly address skewness, gaps and multiple peaks.

#### Skewness

If a skewed distribution is represented purely by two numbers, say the mean and standard deviation, then the representation will be inadequate. As an example the data set below,

9.50	11.9	20.0	33.4	40.1	50.0	12.7	21.0	33.6	40.6
50.0	15.5	26.4	35.4	41.1	50.0	17.7	37.9	41.3	50.0
41.9	50.4	43.0	43.3	43.6	43.7	43.8	44.7	44.9	45.0
45.1	45.2	45.3	46.1	46.5	46.6	47.1	48.0	48.2	
48.5	48.4	48.6	48.7	48.8	48.9	49.4	49.5	49.6	49.8

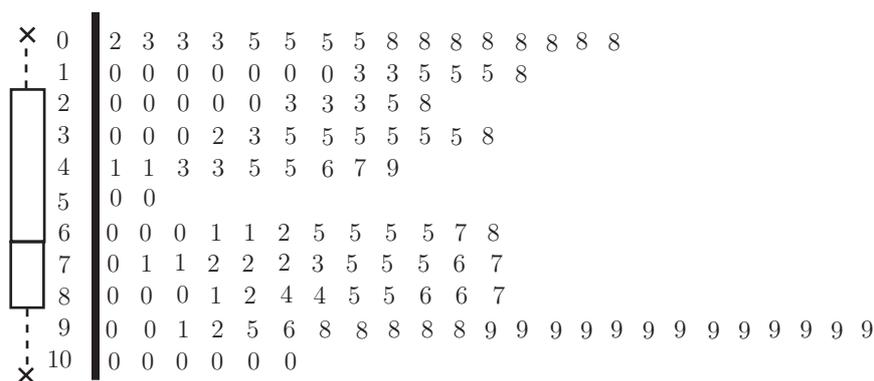
obtained by measuring the current in mA applied to an electronic component under conditions of destructive testing, gives the following values for the mean, standard deviation, median and mid-spread:

$$\bar{x} = 40.72 \quad s = 11.49 \quad \text{median} = 45.05 \quad \text{and} \quad \text{mid-spread} = 9.55$$

The values of  $\bar{x}$  and  $s$  indicate that a lower average current with a greater spread will result in the destruction of the component than that indicated by the median and mid-spread. Clearly, further investigation is necessary to resolve this situation.

## Gaps and Multiple Peaks

Distributions with gaps and multiple peaks can be very difficult to summarise easily. The stem-and-leaf and box-and-whisker plots shown below summarise some 1972 data concerning adult literacy. The leaves are single digit and the range of achievement reached in the field of literacy ranges from 2% to 100%.



The virtual lack of data between 50 and 60 indicates a gap and suggests that we are in fact dealing with two separate distributions which have the following properties:

1. 2% - 50% literacy having **right** skew;
2. 60% - 100% literacy having **left** skew.

Notice that the term skewness refers to the **tail** of a distribution.

The usual summary statistics that you might be tempted to calculate are:

$$\mu = 54 \quad \text{and} \quad \sigma = 34 \quad \text{or} \quad \text{median} = 60 \quad \text{and} \quad \text{mid-spread} = 65$$

In this case, neither set of statistics is of much use since neither set indicates the gap or the skewness. Without visual representation, a single peaked distribution tends to be assumed, this is, of course, opposite to the truth in this case.

The stem-and-leaf plot is more informative than the box-and-whisker plot since it shows the gap.

In practice we would work with the two constituent distributions and attempt to relate the results in a practical way.

## Final Comments on Data Representations

1. You should not rely on summary statistics such as the mean and standard deviation or median and mid-spread alone to represent a data set. Remember that if a distribution has outliers, gaps, skewness or multiple peaks, then shape is probably more important than location and spread.
2. The shape of a distribution is better shown visually than numerically. Remember that a stem-and-leaf diagram retains the data and arranges the data in rank order and that a box-and-whisker plot emphasises the detail contained in the tails of a distribution.

## Exercises

1. The following Used Car prices were taken from a local newspaper. Prices are in £.

150 400 688 795 895 1099 1250 1360 1499 1693  
 250 450 695 890 900 1166 1299 1399 1500 1699  
 300 550 760 895 945 1200 1299 1499 1500 1775  
 350 600 795 895 999 1245 1333 1499 1599 1895

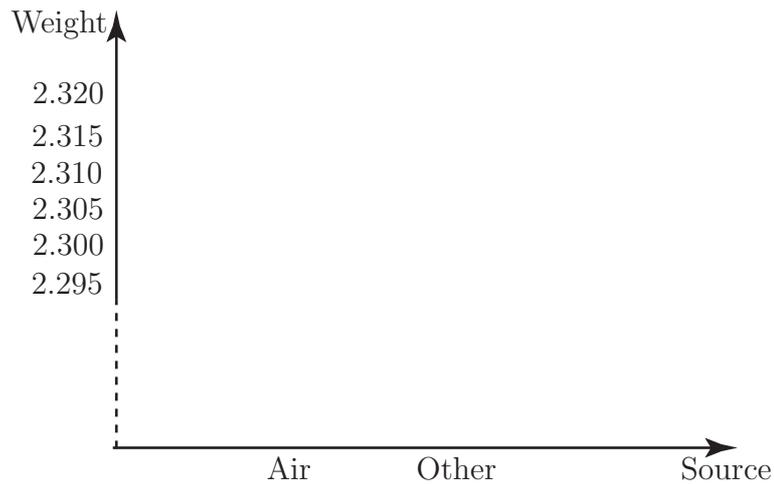
- (a) Represent the data using a stem-and-leaf diagram with two digit leaves.
- (b) Calculate the mean price of the cars on offer using the original data.
- (c) Round the prices to the nearest ten pounds and represent the data using a stem-and-leaf diagram with a single digit leaf.
- (d) Calculate the mean price of the cars on offer using the rounded data.
- (e) Does either mean price give a good indication of the expected price of the cars on offer?
- (f) How would you give a car buyer a good indication of the price he or she can expect to pay for a used car

Comment on the units used in each case with respect to any differences highlighted by the two displays.

2. During the winter of 1893/94 Lord Rayleigh conducted an investigation into the density of nitrogen gas taken from various sources. He had previously found discrepancies between the density of nitrogen obtained by chemical decomposition and nitrogen obtained by removing oxygen from air. Lord Rayleigh's investigations led to the discovery of argon. The raw data obtained during his investigations are given below.

Date	Source	Weight	Date	Source	Weight
29/11/93	NO	2.30143	26/12/93	N <sub>2</sub> O	2.29889
05/12/93	NO	2.29816	28/12/93	N <sub>2</sub> O	2.29940
06/12/93	NO	2.30182	09/01/94	NH <sub>4</sub> NO <sub>2</sub>	2.29849
08/12/93	NO	2.29890	13/01/94	NH <sub>4</sub> NO <sub>2</sub>	2.29889
12/12/93	Air	2.31017	29/01/94	Air	2.31024
14/12/93	Air	2.30986	30/01/94	Air	2.31030
19/12/93	Air	2.31010	01/02/94	Air	2.31028
22/12/93	Air	2.31001			

- (a) Organise the data into a frequency table using the classes 2.29-2.30, 2.30-2.31, 2.31-2.32. Draw the histogram representing the data and comment on any unusual features that you may see.
- (b) Classify the data according to the two sources 'Air' and 'Other'. Order each data set and hence find the median, the hinges and the mid-spreads for each data set. Plot box-and-whisker diagrams for the data on a diagram similar to the one shown below.



Comment on any unusual features that you see. What do the box-and-whisker plots tell you about the nitrogen obtained from the two sources?

### Answers

1.

Stem and Leaf Diagram (2 digit leaves)  
(10s and units)

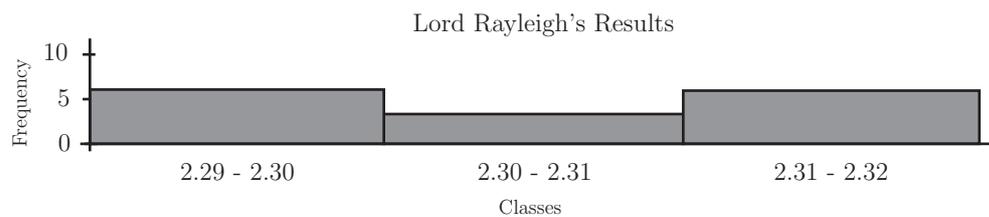
1	50
2	50
3	00,50
4	00,50
5	50
6	00,88,95
7	60,95,95
8	90,95,95,95
9	00,45,99
10	99
11	66
12	00,45,50,99,99
13	33,60,99
14	99,99,99
15	00,00,99
16	93,99
17	75
18	95

Stem and Leaf Diagram (single digit leaves)  
(10s)

1	5
2	5
3	0,5
4	0,5
5	5
6	0,9
7	0,6
8	0,0,9
9	0,0,0,0
10	0
11	0,7
12	0,5,5
13	0,0,3,6
14	0
15	0,0,0,0,0
16	0,9
17	0,8
18	
19	0

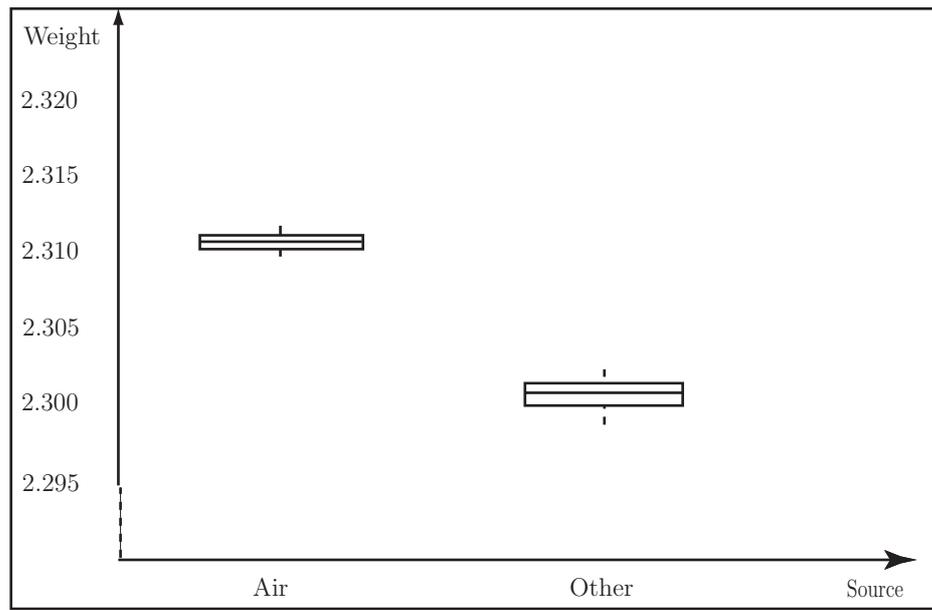
The two-leaf diagram gives a better indication of the spread of data as no accuracy is lost. The single digit diagram is simple and gives *ball-park* figures but loses accuracy due to the rounding process.

2. (a)



The lowest class is obtained entirely from non-air sources, the highest class is obtained entirely from air.

(b)



Comment. Box-and-whisker plot tells us that some other element is present in Air which is responsible for the additional weight. This *additional* element subsequently proved to be the inert gas argon.