

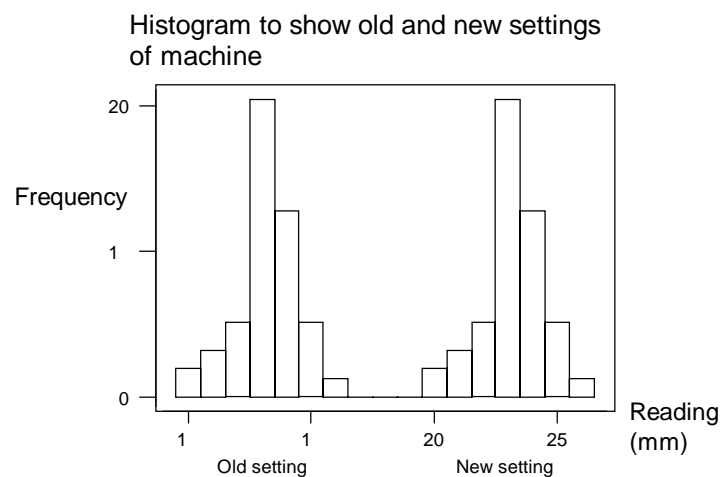
## 1.5 NUMERICAL REPRESENTATION OF DATA (Sample Statistics)

As well as displaying data **graphically** we will often wish to summarise it **numerically** particularly if we wish to compare two or more data sets. There are three fundamental properties of any data set in which we are interested:

- *shape*
- *location (or average)*
- *spread (or variability)*

### 1.5.1. Measures of Location

A change in the position of a histogram corresponds to a change in the **location** or position of the data set (distribution).



There are a number of measures of location (or average) that are commonly used. These are the **mean**, the **median** and the **mode**.

#### (a) Mean

Probably the most common measure of location is the *mean* of the data set, denoted by  $\bar{X}$ .

The mean is obtained by adding together all the values, or observations, and dividing by the number of values.

Let  $X_i$  = *i*th observation  
 $n$  = number of observations

Then

$$\bar{X} = \frac{\sum X_i}{n}$$

**Example**

Suppose a machine has been set up to cut pieces of steel to a certain length. Ten pieces of steel in succession are removed from the production line, and their lengths measured (mm) as follows:

143	141	140	139	142	141	142	143	139	142
$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$

Calculate the mean length of the steel pieces sampled.

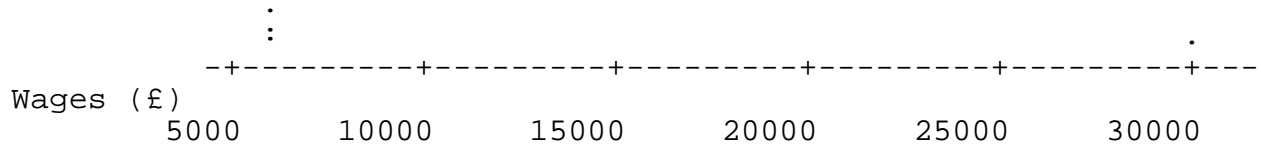
Then

$$\begin{aligned} \text{mean} = \bar{X} &= \frac{\sum X_i}{n} \\ &= \frac{X_1 + X_2 + X_3 + \dots + X_{10}}{10} \\ &= \frac{143 + 141 + 140 + \dots + 142}{10} \\ &= 141.2 \text{ mm} \end{aligned}$$

One problem with the mean as a measure of location is that it is very sensitive to outlying (extreme) values. In these situations, if the mean is used as a 'typical value' it can give a false impression of the data. For example, a small firm consists of:

- 1 manager earning £30,000 per year
- 3 manual workers each earning £6,000 per year.

If we plot this data on a dotplot it looks like :



i.e. we have an extreme value of £30,000.

Calculating the mean

$$\begin{aligned}\bar{X} &= \frac{\sum X_i}{n} = \frac{6000 + 6000 + 6000 + 30000}{4} \\ &= \text{£}12000\end{aligned}$$

However, this mean salary is not representative since it's double what most people earn.

An alternative measure of location which is much less sensitive to extreme values (**robust**) is the **median**.

(b) **Median**

The median is defined to be the value which divides the data into two equal parts, i.e. the 'middle' value. Half the values are above it and half are below it.

To obtain the median

- (i) order the data in ascending order.
- (ii) For  $n$  items in the data set, the median is the

$$\left(\frac{n+1}{2}\right)\text{th observation.}$$

**Example**

Consider the data set shown in the previous example. Find the median value.

First, order the data.

139	139	140	141	141	142	142	142	143	143
X <sub>4</sub>	X <sub>9</sub>	X <sub>3</sub>	X <sub>2</sub>	X <sub>6</sub>	X <sub>5</sub>	X <sub>7</sub>	X <sub>10</sub>	X <sub>1</sub>	X <sub>8</sub>

Then, since we have 10 observations (data points) the median value is the

$$\left(\frac{10+1}{2}\right) \text{th observation} = (11/2) = 5.5\text{th observation.}$$

Now: 5th observation = 141 (X<sub>6</sub>)  
6th observation = 142 (X<sub>5</sub>)

So 5.5th observation = mean of 5th and 6th observations.  
i.e. Median = 141.5

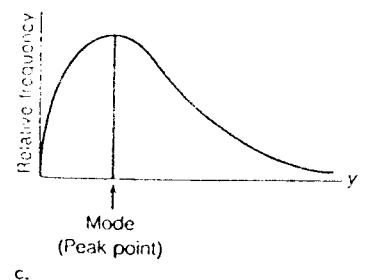
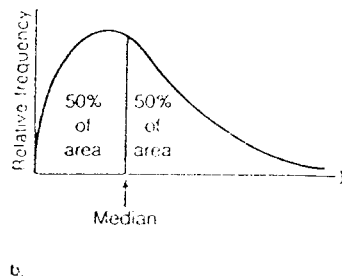
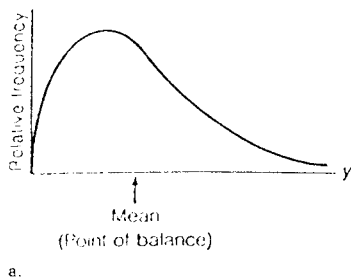
In this case, the median is very similar to the mean (=141.2) though this will not always be so.

(c) **Mode**

The mode of a set of n measurements X<sub>1</sub>, X<sub>2</sub>, .....X<sub>n</sub> is the value of X that occurs with the greatest frequency. (i.e. the most popular or common value)

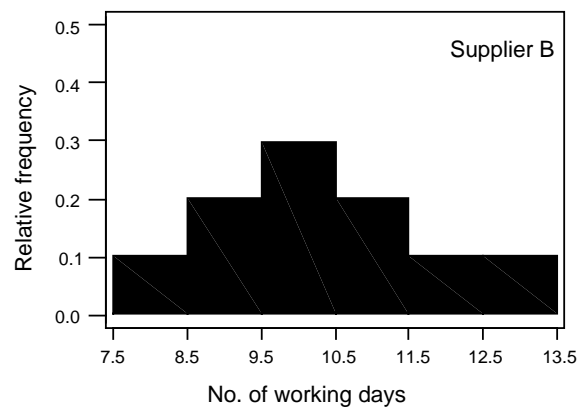
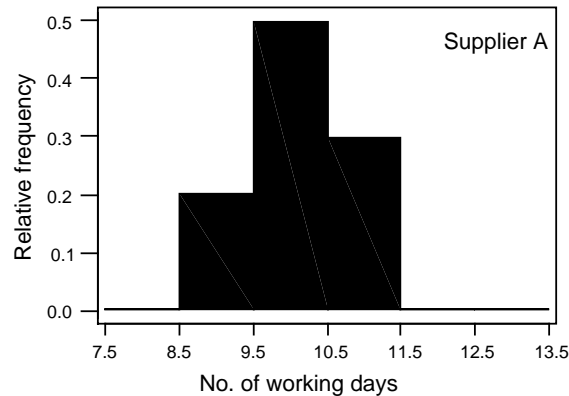
It is rarely used as a measure of location unless the shape of the distribution is bimodal in which case no single measure of location gives a reasonable description.

The three measures of location can be represented diagrammatically as follows:



## 1.5.2. Measures of Variation

Assume that you are a purchasing agent for a large manufacturing firm and that you regularly place orders with two different suppliers. After several months of operation you find that the mean number of days required to fill orders is averaging around 10 days for both suppliers. However, histograms based on historical data are shown below.



Which supplier would you prefer to deal with?  
{Answer: Supplier A, as times are less variable}

Although the mean supply time is the same for both suppliers we can see that there is more variation in the Supplier B delivery times. We need to be able to measure this variation so that comparisons can be made.

Commonly used measures of variation are:

- *Range*
- *Interquartile Range*
- *Standard Deviation*

(a) **Range**

The simplest measure of the spread or variation of a set of data is the range,  $R$ , which is defined as the difference between the largest and the smallest value.

Consider the data given in a previous example

i.e. 10 datum heights (in mm):

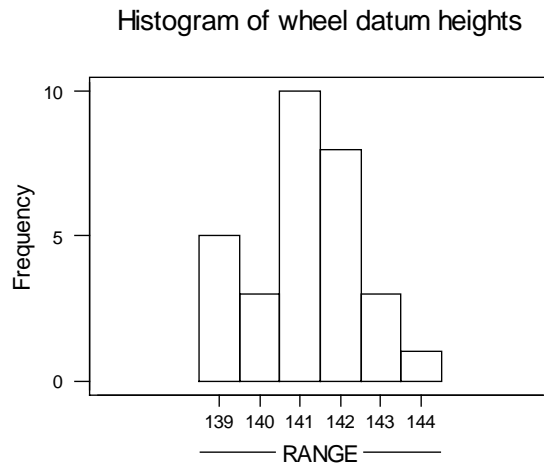
143 141 140 139 142 141 142 143 139 142

$$\begin{aligned} \text{Then Range} &= R = \text{Maximum} - \text{Minimum} \\ &= 143 - 139 \\ &= 4 \text{ mm} \end{aligned}$$

If the data are arranged in a histogram, the range corresponds to the width of the base. The following is based on a sample of 30 turbo wheels - datum heights measured in cm.

Turbo wheel height (cm)	Frequency	Relative frequency
138.5 - (139.5)	5	$\frac{5}{30} = 0.167$
139.5 - (140.5)	3	0.1
140.5 - (141.5)	10	0.333
141.5 - (142.5)	8	0.267
142.5 - (143.5)	3	0.1
143.5 - (144.5)	1	0.033
	30	1.000

The corresponding histogram is then :



There are TWO problems with the Range as a measure of variation:

1. Although the range is a simple-to-use measure of variation, it only uses the two extremes of the data set (i.e. the maximum and minimum values). Thus, it is significantly affected by extreme (outlying) values within the data.
2. In addition, it measures spread on a scale that is dependent of sample size. (i.e. the range of a sample of 10 is not comparable with the range of a sample of size 100.)

A range-based measure which is not so affected by extremes is the:

(b) **Interquartile Range**

The quartiles of a set of data are the values which divide the ordered data set into four equal parts. They are denoted by

$Q_1$  = first quartile (or lower quartile)

$Q_2$  = second quartile (or median)

$Q_3$  = third quartile (or upper quartile)

For example, suppose we have 11 observations (data points) as shown

11, 6, 4, 10, 1, 8, 5, 9, 2, 3, 7

We first order the data, (i.e. put the values into ascending numerical order) so that we get

1	2	3	4	5	6	7	8	9	10	11
		$Q_1$			$Q_2$			$Q_3$		

Then  $Q_1$ ,  $Q_2$  and  $Q_3$  are such that the data is divided equally into quarters as shown.

In fact, for  $n$  observations,

$Q_1$  is the  $\left(\frac{n+1}{4}\right)$ th data point

$Q_2$  is the  $\left(\frac{n+1}{2}\right)$ th data point

$Q_3$  is the  $3\left(\frac{n+1}{4}\right)$ th data point

Then the Interquartile range (IQR) is defined to be

$$\text{IQR} = Q_3 - Q_1$$

It is the range of the middle half of the data.

### Example

The value (in £ thousands) of the items sold by Cornwood electrical company in each month for 1991 was as follows:

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
Value	51	53	54	51	51	55	61	62	63	61	59	75

First order the data:

51    51    51    53    54    55    59    61    61    62    63    75

$Q_1$ , the first quartile, is the point taken to be  $\frac{n+1}{4}$  of the way along the ordered data set. (Remember  $n$  represents the number of data points)

$$\text{i.e. } Q_1 \text{ is the } \frac{(n+1)}{4} = \frac{(12+1)}{4} = \frac{13}{4} = 3.25\text{th data point}$$

Thus  $Q_1$  is one quarter (0.25) of the distance between the third largest and the fourth largest observation i.e. one quarter of the way between 51 and 53.

Hence:  $Q_1 = 51.5$

$Q_3$  is the point taken to be  $\frac{3(n+1)}{4}$  of the way along the ordered data set.

$$\text{Hence: } \frac{3(n+1)}{4} = \frac{3(12+1)}{4} = \frac{3 \times 13}{4} = 9.75\text{th data point}$$

Thus  $Q_3$  is three quarters (0.75) of the distance between the ninth largest and tenth largest observation, (i.e. 61 and 62).

Hence:  $Q_3 = 61.75$

Hence the Interquartile range is:

$$Q_3 - Q_1 = 61.75 - 51.5 = 10.25 \text{ or } \pounds 10,250.$$

The Interquartile range is less affected by the extreme values than the range. However it still only uses the relative position of two of our data points.

### **Quantiles and Percentiles**

The idea of quartiles (which divide a set into 4 equal parts) can be generalised to statistics which divide the data into any number of equal parts, called **quantiles**. Of particular importance are the **percentiles** which divide the data into 100 equal parts.

For example:

1 <sup>st</sup> percentile, $p_1$ ,	has 1% of values below it,	99% above it
5 <sup>th</sup> percentile, $p_5$ ,	has 5% of values below it,	95% above it
50 <sup>th</sup> percentile, $p_{50}$ ,	has 50% of values below it,	50% above it
95 <sup>th</sup> percentile, $p_{95}$ ,	has 95% of values below it,	5% above it
99 <sup>th</sup> percentile, $p_{99}$ ,	has 99% of values below it,	1% above it

Percentiles are particularly useful when dealing with probability distributions.

### (c) **Standard Deviation**

An alternative measure of variation, which uses all the available information (i.e. all of the observations, or data points) is the **standard deviation, s**. It is based on the idea of measuring how far, on average, the observations are from the centre of the data.

To illustrate the principle, consider the following lengths of offcuts (mm) of five rods:

5      3      7      1      4

The mean length  $\bar{X} = 4\text{mm}$ .

	Reading (mm) X	Deviation from mean (X - $\bar{X}$ )	Squared Deviations (X - $\bar{X}$ ) <sup>2</sup>
	5	1	1
	3	-1	1
	7	3	9
	1	-3	9
	4	0	0
TOTALS	20	0	20

For technical reasons, to average the squared deviations we divide by  $n - 1$  rather than  $n$  (this gives an unbiased estimate of the equivalent population characteristic).

Then

$$\frac{\sum (X - \bar{X})^2}{n - 1} = \frac{20}{4} = 5$$

This is called the **variance** and is denoted  $s^2$ .

The variance is measured in the square of the units of the original data ( $s^2 = 5 \text{ mm}^2$  in the above) while all the other summary statistics we have considered are in the same units as the data. Thus a better measure for practical use is the square root of the variance called the **standard deviation** defined by:

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} \quad (= \sqrt{5} = 2.236 \text{ mm in above})$$

The SD is zero if there is no variation in the data. The larger the SD, the more variation there is.

For hand calculations it is more convenient to use the following formula:

$$s = \sqrt{\frac{1}{n - 1} \left\{ \sum X^2 - \frac{(\sum X)^2}{n} \right\}}$$

as follows

X = offcut length	X <sup>2</sup>
5	25
3	9
7	49
1	1
4	16

Then  $\Sigma X = 5+3+7+1+4 = 20$

$$\Sigma X^2 = 25+9+49+1+16 = 100$$

and  $s = \sqrt{\frac{\{100 - (20)^2 / 5\}}{4}}$

$$= \sqrt{\frac{20}{4}}$$

$$= \sqrt{5}$$

$$= 2.236$$

Thus the 5 offcut measurements have a mean of 4mm and a standard deviation of 2.236mm.

### Using a Calculator

Most calculators have a facility to calculate means and standard deviations automatically. The exact procedure depends on the model but generally it involves

- (i) **Get into SD or STAT mode.** SD will appear on the display
- (ii) **Clear memories** by pressing INV or SHIFT then AC.
- (iii) **Enter data.** Enter each value followed by the DATA key. (M+ on some calculators.)
- (iv) **Push  $\bar{X}$  button** for mean and  $\sigma_{n-1}$  or s button for the standard deviation. (See section 1.5.4 comment on this)

Repeat the offcut example using your calculator facilities.

## Comparison of range and standard deviation

### Example

The examples below illustrate how the range and standard deviation compare for various samples.

		MEAN $\bar{X}$	RANGE R	S.D. S
<u>Data Set 1</u> 5 5 5 5 5	$\frac{x}{0 \quad 10}$	5	0	0
<u>Data Set 2</u> 4 5 5 5 6	$\frac{\quad}{0 \quad 10}$	5	2	0.7
<u>Data Set 3</u> 1 5 5 5 9	$\frac{\quad}{0 \quad 10}$	5	8	2.8
<u>Data Set 4</u> 1 3 5 7 9	$\frac{\quad}{0 \quad 10}$	5	8	3.2
<u>Data Set 5</u> 1 3 3 9 9	$\frac{\quad}{0 \quad 10}$	5	8	3.7
<u>Data Set 6</u> 8 9 9 9 10	$\frac{\quad}{0 \quad 10}$	9	2	0.7

- Note:** 1 The Range does not reflect any difference in the level of variation for data sets 3, 4 and 5. Thus, it is rather 'crude'.
- 2 Comparing data sets 2 and 6, we see that just shifting the data 'along the scale' (all values increased by 4 in this case) has no effect on the S.D.

This is useful, for example, if required to calculate the mean and standard deviation of:

21876            21875            21872            21871            21876

What is the easiest approach?

Since 21871 is the smallest number, let this be equivalent to zero.

Then  $21872 \rightarrow 21872 - 21871 = 1$   
 $21875 \rightarrow 21875 - 21871 = 4$   
 $21876 \rightarrow 5$

i.e. Data set is equivalent to

5      4      1      0      5

for the purposes of calculating the SD.

Then  $\Sigma X = 5 + 4 + 1 + 0 + 5 = 15$

$$\Sigma X^2 = 25 + 16 + 1 + 0 + 25 = 67$$

and

$$s = \sqrt{\frac{1}{4} \left\{ 67 - \frac{(15)^2}{5} \right\}}$$

$$= \sqrt{\frac{1}{4} \{67 - 45\}}$$

$$= \sqrt{\frac{1}{4} (22)}$$

$$= \sqrt{5.5}$$

$$= 2.345\text{mm}$$

### 1.5.3 Choice of Measures

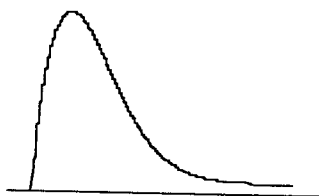
To summarise a set of data we require one measure of location and one measure of variation. The mean and SD are the preferred measures because they use all the available data. However, they are affected by extreme values. Thus, for descriptive purposes, use as follows:



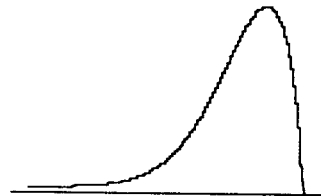
(Roughly symmetrical)



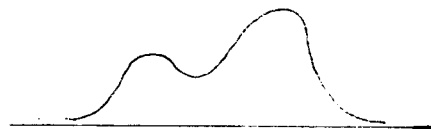
*Mean and Standard Deviation*



(Very skew)



*Median and IQR*



(Bimodal)

*Modes and Range*

An important exception to the above pairings is used in Statistical Process Control (SPC) procedures for monitoring some quality characteristic. Here, the mean and range are often used (in the familiar  $\bar{X} - R$  charts) because they are easiest to calculate and interpret.

Finally, note that additional summary statistics are often produced by computer packages. The Excel spreadsheet, for example, also gives measure of skewness and measures of kurtosis (or flatness), but these are rarely used in practice.

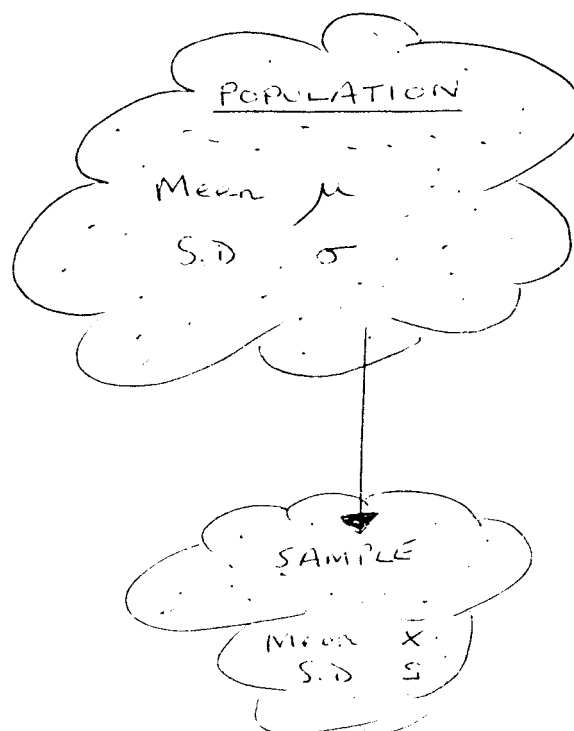
### 1.5.4 Parameters and Statistics

Distinctions between populations and samples should also be made when dealing with descriptive measures.

*Quantities describing features of samples (such as  $\bar{X}$  and  $s$ ) are called **sample statistics**.*

*Quantities describing features of populations are called **population parameters**.*

Important examples of parameters are the true population mean, denoted  $\mu$ , and the true population standard deviation, denoted  $\sigma$ . Then, for example,  $\bar{X}$  and  $s$  can be used to estimate the usually unknown  $\mu$  and  $\sigma$ .



- Population parameters are **FIXED** constants.
- Sample statistics **VARY** from sample to sample.

We are only interested in sample statistics in terms of what they tell us about the unknown population parameters.

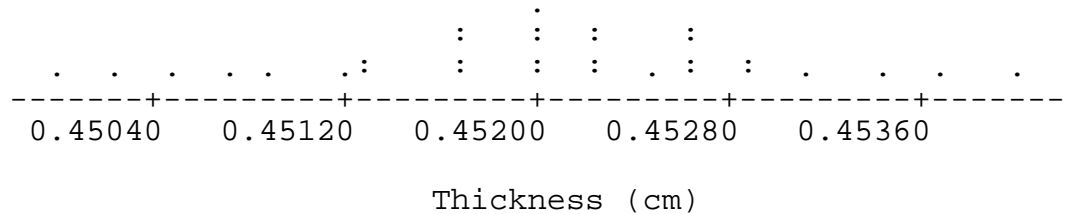
*(Final Note: Many calculators, including Casio, use confused notation. The mean and SD buttons are marked  $\bar{X}$  and  $\sigma$ . This is simply wrong and should be avoided).*

## 1.6. STRATIFICATION

The scope and power of diagrams can be increased using the idea of stratification. If the data come from different known sources (e.g. machines, departments, individuals), this involves plotting for each source separately. Similarly, summary statistics can be calculated for each source separately.

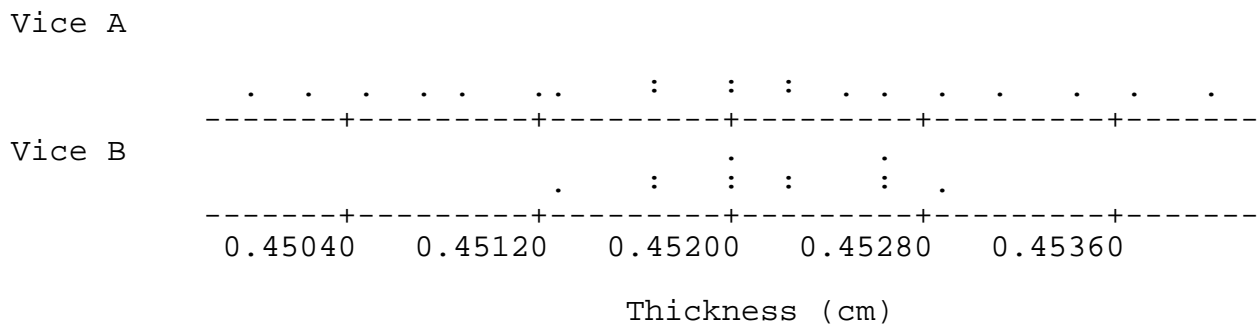
### Example 1

Consider data available on the flange thickness of a machined part. This is continuous data. Then the Dot plot (overall) is:



Suppose that the flanges are actually formed on two different vices. If we separate the information into 2 sets of data, one for each vice, i.e. Data is *stratified by vice*.

Then the Dot plot (stratified by vice) is:

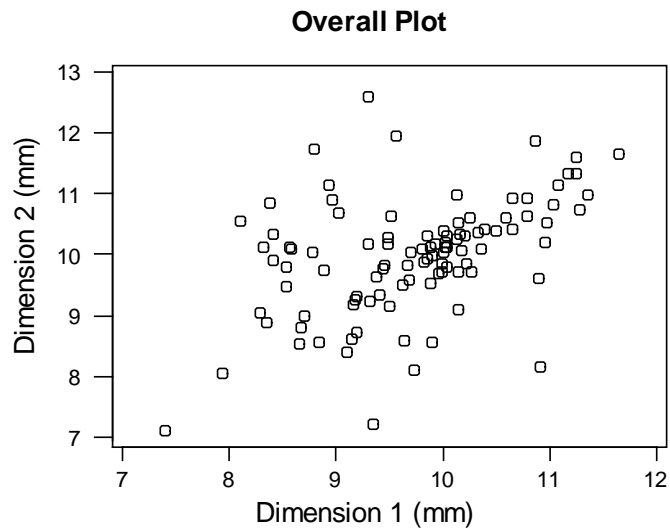


### Interpretation:

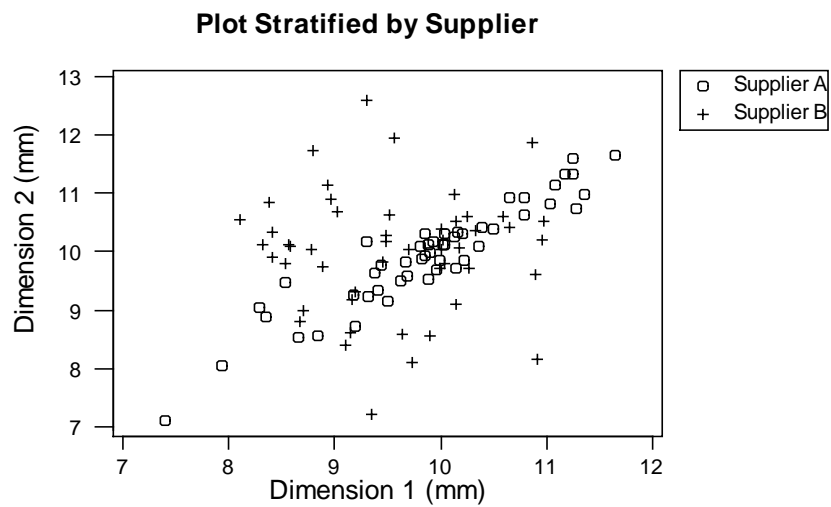
Both vices produce flanges with approximately the same 'average' or middle value (i.e. 0.452cm). However, there is more 'variation' or 'spread' in the flanges produced by Vice A (i.e. flanges vary from 0.450 to 0.454cm). Vice B is the better of the two.

**Example 2**

A component has two critical dimensions which should be related (if one dimension is relatively large then so should the other one be). A scatter plot of a sample of components suggests there is a problem here. (This is example 2, section 1.4.5.)



However, these components come from two suppliers and the scatter plot stratified by supplier suggests that the problem really only lies with supplier B.



## 1.7 COMPARISON OF DATA SETS

We have previously shown how various measures of location (typical values) and variation (spread) can be calculated for a particular data set. Whilst these are useful for a single set of data, it is more usual to use the measures for comparing two or more data sets.

For a full comparison we need to compare

- **shape**
- **location**
- **spread**

and to identify if there are any **outliers** (odd extreme values) present. If there are any, find the reason for their 'extremeness' and remove them from the data set.

### 1.7.1. Guidelines for comparison

(i) Draw a suitable diagram to illustrate the data.

NB. In making comparisons between different data sets, we must be sure that we compare 'like with like'. Thus, if two (or more) sets contain very different numbers of observations, comparing actual **numbers** in classes will be inappropriate - **relative frequencies** must then be used.

(ii) Choose a suitable measure of location (this will show where the data is concentrated) and choose a suitable measure of variation (this will show the spread).

(iii) Calculate (or use a statistical package to obtain) the measures.

(iv) Summarise the measures, for example, as a table.

(v) Draw conclusions from the summary.

### Example

A car manufacturer has a limited range of small cars that have proved very successful. They would like to add a new edition to the range and in order to launch this in the most potentially profitable area they record the daily sales of all the other models (in total) over a three month period for a number of garages in an area of each of Devon, Cornwall and Somerset. The tabulated results of their recordings are as follows:

Daily Sales ('0000's)	Area 1	Area 2	Area 3
2 - (4)	2	1	5
4 - (6)	2	2	5
6 - (8)	8	9	6
8 - (10)	17	25	14
10 - (12)	38	25	10
12 - (14)	28	19	8
14 - (16)	28	13	6
16 - (18)	18	6	5
18 - (20)	10	0	3
20 - (22)	3	1	2
22 - (24)	0	0	1
Total	154	101	65

By comparison of the sales in the three areas, make a recommendation to the sales manager as to the best area in which to make the initial launch of the new edition.

### Solution

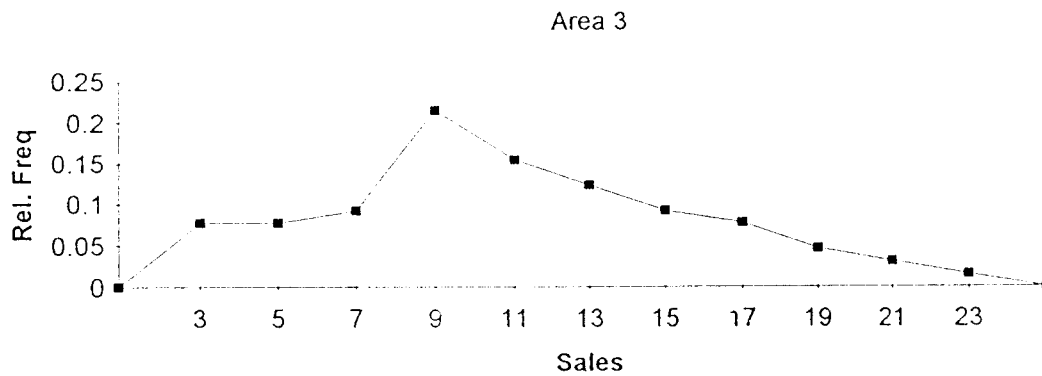
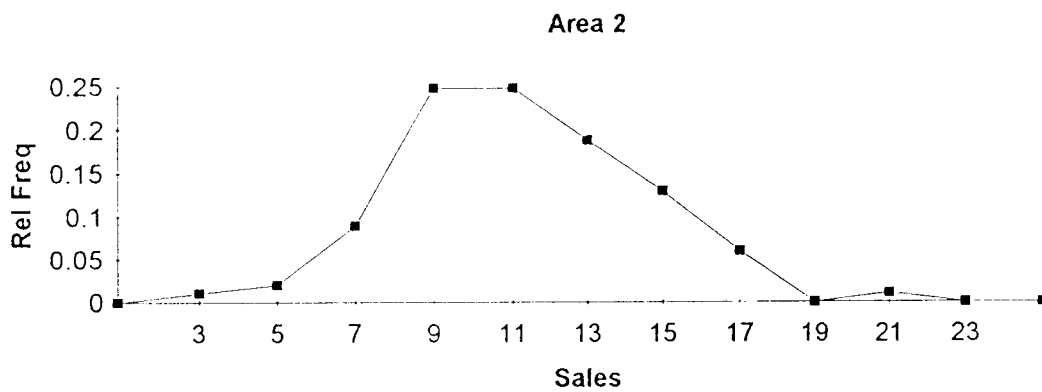
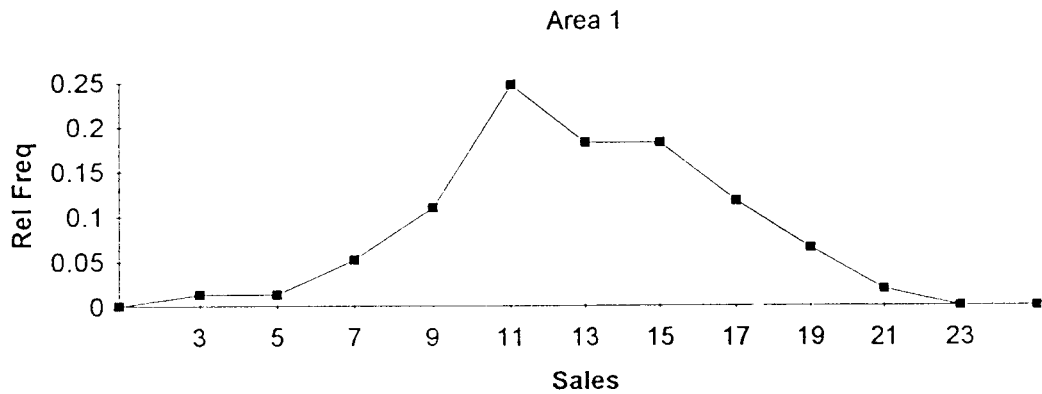
1. **PLOT** Since there are different numbers of observations in each area, we must use the relative frequencies.

Class	Area 1	
	Freq.	Rel. Freq
2 - (4)	2	0.013
4 - (6)	2	0.013
6 - (8)	8	0.052
8 - (10)	17	0.110
10-(12)	38	0.247
12-(14)	28	0.182
14-(16)	28	0.182
16-(18)	18	0.117
18-(20)	10	0.065
20-(22)	3	0.019
22-(24)	0	0
	154	1

A Summary of Output for all 3 Areas

Class	Area 1		Area 2		Area 3	
	Freq.	Rel. Freq	Freq.	Rel. Freq	Freq.	Rel. Freq
		Cum. Freq		Cum. Freq		Cum. Freq
2 - (4)	2	0.013	1	0.010	5	0.077
4 - (6)	2	0.013	2	0.020	5	0.077
6 - (8)	8	0.052	9	0.089	6	0.092
8 - (10)	17	0.110	25	0.248	14	0.215
10-(12)	38	0.247	25	0.248	10	0.154
12-(14)	28	0.182	19	0.188	8	0.123
14-(16)	28	0.182	13	0.129	6	0.092
16-(18)	18	0.117	6	0.059	5	0.077
18-(20)	10	0.065	0	0.000	3	0.046
20-(22)	3	0.019	1	0.010	2	0.031
22-(24)	0	0	0	0	1	0.015
	154	1	101	1	65	1

A suitable diagram requires the relative frequencies for each data set.



**NOTE:** These diagrams are placed one above the other for ease of comparison. However, it is acceptable to draw all three distributions on the same set of axes provided each is shaded or coloured differently. In this case, frequency polygons are preferred to histograms.

**2. CHOOSE MEASURES** From the relative frequency polygons we can see that each is reasonably symmetric; area 3 is located to the left of the areas 1 and 2; all three have similar spreads. We choose the mean and standard deviation as the measures of location and variation respectively.

**3 & 4. CALCULATE AND SUMMARISE**

	Area 1	Area 2	Area 3
Sample size	154	101	65
Mean	128641	112579	86200
Standard Deviation	35285.4	31958.6	33897.5

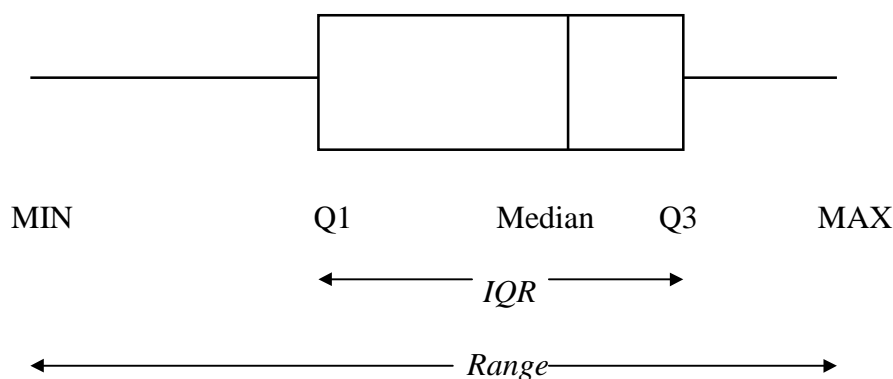
**5. INTERPRET** From the summary above, we can see that the standard deviations (spread) are indeed very similar, although area 2 is slightly less spread out than areas 1 or 3. Area 1 has the highest mean sales, £128,641, which is 14% higher than area 2, the next highest. If sales are the only factor to be considered in this situation, then we can recommend to the sales manager that Area 1 be chosen for the launch of the new car.

**1.7.2. Boxplots**

A graph for summarising the features of a set of data that is particularly useful when comparing data sets is a **boxplot**. This is based on a 'five-number' summary of the data:

- Minimum value
- Lower quartile ( $Q_1$ )
- Median ( $Q_2$ )
- Upper quartile ( $Q_3$ )
- Maximum value

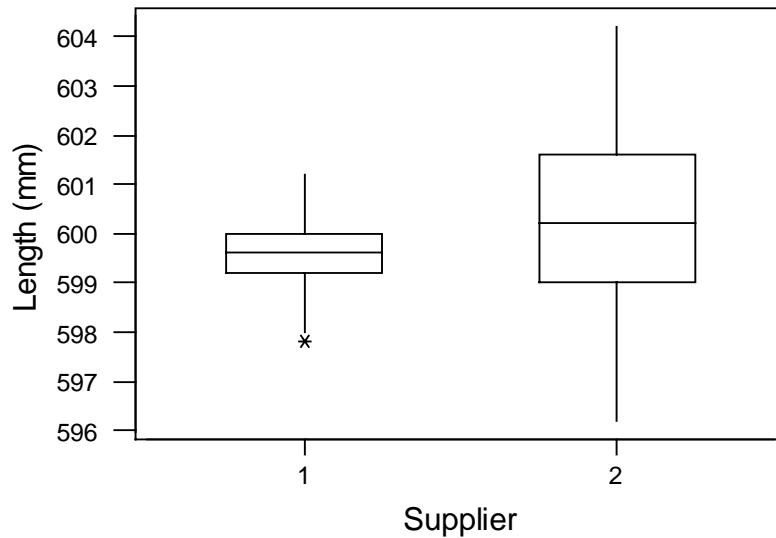
These are represented as:



The width of the central box is arbitrary. Many packages (including Minitab) indicate extreme outlying values as dots or crosses as in the following examples.

**Example 1**

Comparison of camshaft lengths from 2 suppliers



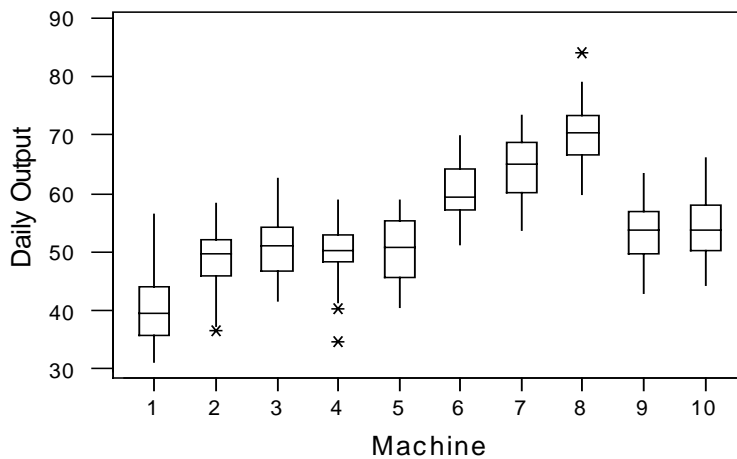
***Interpretation***

The plots clearly show that the average (median) lengths are approximately the same for both suppliers but that there is much greater variation in length of camshafts coming from supplier 2. It can also be seen that both distributions of length are symmetrical. The single outlier flagged from supplier 1 may need further investigation but it doesn't seem to be that far away from the rest of the data.

**Example 2**

Boxplots are particularly useful when comparing many data sets as in the following :

Comparison of performance of 10 machines



## 1.8 REVIEW

This section has been concerned with summarising the information contained in a sample.

Data can be either

- Quantitative (discrete/continuous)
- Qualitative (categorical/attribute)

Quantitative (numerical) data can be summarised by

- Tables (especially frequency distributions)
- Diagrams (especially histograms)
- Statistics (especially mean and SD/median and IQR)

Frequency distributions and histograms describe the shape observed in the sample and suggest what the corresponding probability distributions are in the population.

The mean and SD ( $\bar{X}$  and  $s$ ), calculated from a sample, define more precisely certain features of the data;  $\bar{X}$  measures location (or 'average'),  $s$  measures variation or spread. They estimate the corresponding 'true' values, i.e. the population parameters  $\mu$  and  $\sigma$ .

We have also considered the use of stratification to 'break down' a set of data for more meaningful comparisons.