EDEXCEL NATIONAL CERTIFICATE

UNIT 4 – MATHEMATICS FOR TECHNICIANS
OUTCOME 3

STATISTICAL METHODS

Use **statistical methods** to gather, manipulate and display scientific and engineering data

CONTENTS

Data manipulation**:** gathering and collation of data from varying sources; grouped and non-grouped data, frequency; graphical representation of statistical data, using bar charts, pie-charts and histograms

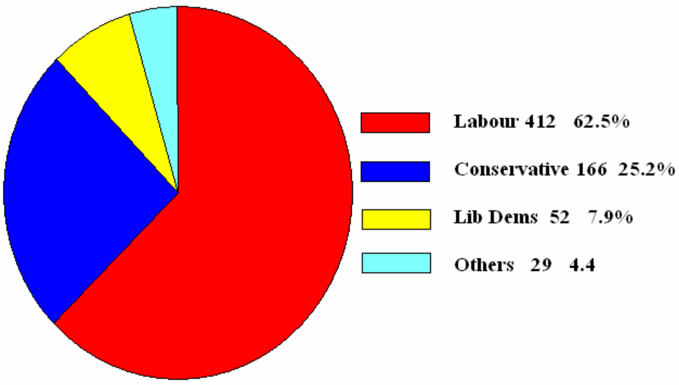
Statistical measurement**:** measures of central tendency from group data – mean, median, mode, standard deviation and variance; use of calculator to manipulate statistical functions.

On completion of this tutorial you should be able to do the following.

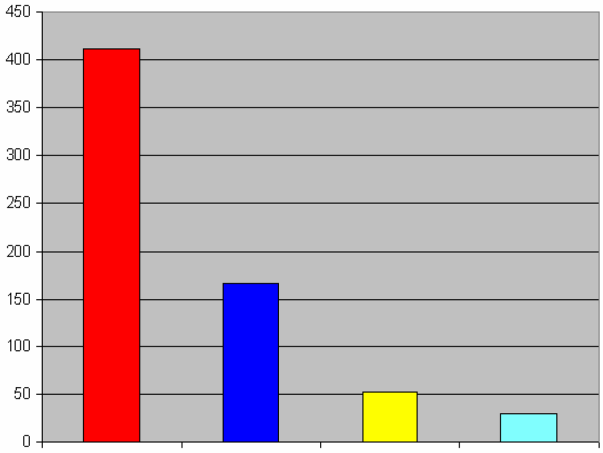
- Explain the use of raw data.
- Present data as frequency polygons, bar graphs and histograms.
- Explain and find the mean and median values.
- Explain and plot ogives.
- Explain and find quartiles.
- Explain and find the standard deviation and variance for grouped and ungrouped data.

1. INTRODUCTION

Statistics are used to help us analyse and understand the performance and trends in various areas of work. These might be financial trends, things to do with the population or things to do with manufacturing. Often we wish to present information visually with easily understood graphics and so a variety of graphs and charts are used for this purpose. Here is an example of a PIE CHART and a BAR CHART showing the number of parliamentary seats won by main political parties at the 2001 British general election.



PIE CHART



BAR CHART

The number of MPs elected must be a round or whole number so the raw data is exact. When presenting other forms of data this is not the case as explained in the following section.

2. RAW DATA

Let’s use an example to get started. Consider a set of statistics compiled for the height of children of age 10. First we would compile a table of heights. This would be the raw data. Note that the larger the sample we take, the more meaningful the results will be. Suppose we measure the heights to an accuracy of 0.01 m. This is the table of raw data for 10 year old children

Sample	1	2	3	4	5	6	7	8	9	10	11	12
Height (m)	1.45	1.56	1.37	1.44	1.32	1.42	1.55	1.29	1.37	1.49	1.47	1.34
Sample	13	14	15	16	17							
Height (m)	1.56	1.28	1.35	1.62	1.46							

3. RANKED DATA

If the raw data is rearranged from shortest to tallest we have the ranked data.

Sample number	1	2	3	4	5	6	7	8	9	10	11	12	13
Height (m)	1.28	1.29	1.32	1.34	1.35	1.37	1.37	1.44	1.44	1.46	1.46	1.47	1.49

Sample number	14	15	16	17	Total
Height (m)	1.55	1.56	1.56	1.62	24.34

Data presented in this form is also called **DISCRETE DATA** because we jump from one value to another in steps, in this case steps of 0.01 m.

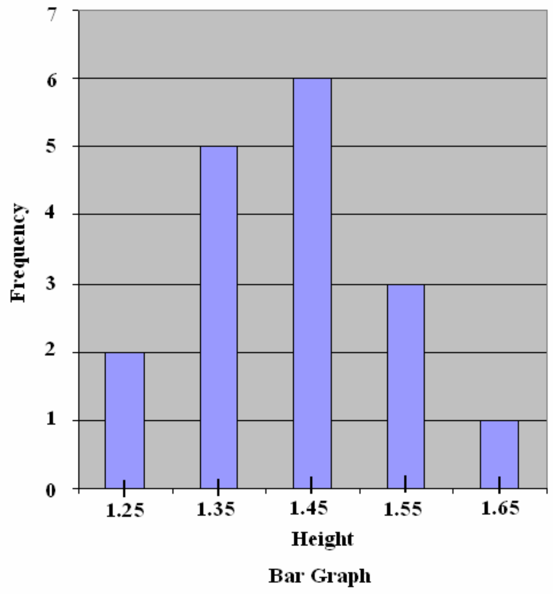
4. CLASS or BANDS

If we used exact measurements of heights, it is unlikely we would find two children exactly the same height so we round off the values. This causes problems as we shall see. When handling large numbers of samples, we end up with huge lists of data so to simplify the table we create bands or classes within which the rounded measurements fall. The more we round off the values, the more likely it becomes that we will find more than one in a given class. The number of children within each class is the ***frequency***. Next we would have to go through the laborious task of counting how many there are in each class. If we found a child with a height exactly on the edge of the class edge, we might decide to allocate a half to each class on either side resulting in frequency values that are not whole numbers. The result is a **FREQUENCY DISTRIBUTION TABLE.**

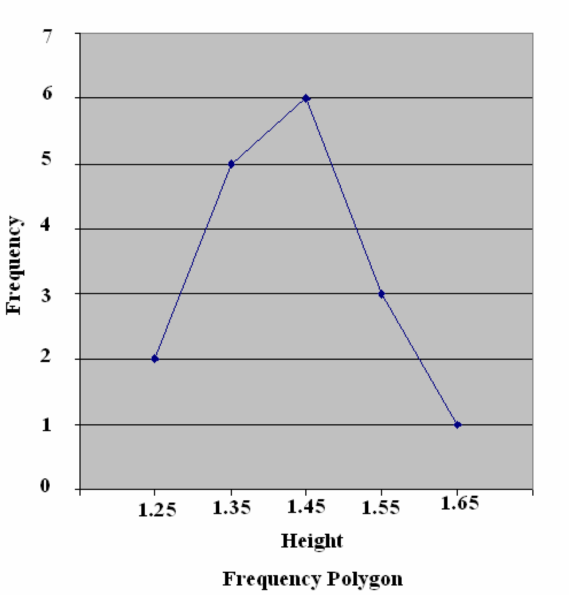
Height	1.2 - 1.3	1.3 - 1.4	1.4 - 1.5	1.5 - 1.6	1.6 - 1.7
Mid Point or MARK	1.25	1.35	1.45	1.55	1.65
Freq.	2	5	6	3	1

5. GRAPHS

If we plot frequency vertically against height horizontally, we get a frequency distribution graph and this can be drawn in different ways. The values plotted are the mid point values called the MARK.



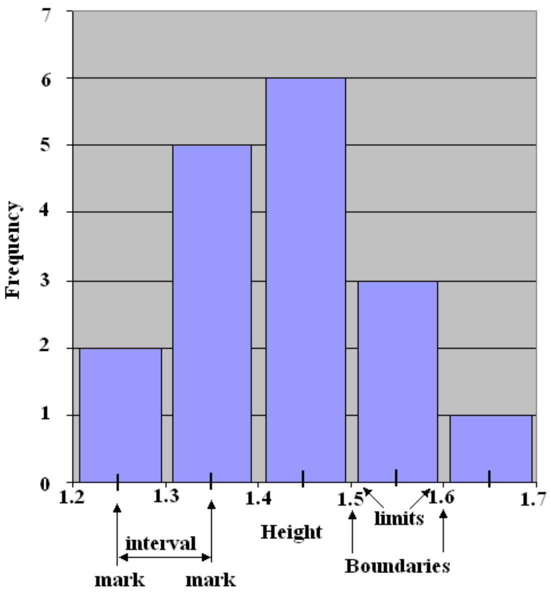
Bar Graph



Frequency Polygon

These plots simply tell us the numbers in each class by the mark. If we want to illustrate the width of the band we use a **HISTOGRAM**. Notice that the mid point of each band is the **MARK**. The boxes are drawn between the **CLASS LIMITS**. Because the heights were rounded off to 0.01 m the limits are 0.005 either side of the **CLASS BOUNDARY** and the class boundary is the exact dividing line between each class. The width of the band from mark to mark or boundary to boundary is the **CLASS INTERVAL.** In this example the interval is 0.1.

Notice that the points are drawn for the middle of the band.



6. MEAN

This is one of the more common statistics you will see and it's easy to compute. All you have to do is **add** up all the values in a set of data and then **divide** that sum by the number of values in the dataset. For our example, let the height be represented by the variable x and the frequency be f.

Sample number	1	2	3	4	5	6	7	8	9	10	11	12	13
Height (m)	1.45	1.56	1.37	1.44	1.32	1.42	1.55	1.29	1.37	1.49	1.47	1.34	1.56

Sample number	14	15	16	17	Total
Height (m)	1.28	1.35	1.62	1.46	24.34

The mean value is denoted $\bar{x}$ and $\bar{x} = 24.34/17 = 1.432$ m

We can do this a bit more simply using the frequency distribution table.

x (mid pt)	1.25	1.35	1.45	1.55	1.65	
f.	2	5	6	3	1	total
f x	2.5	6.75	8.7	4.65	1.65	24.25

The mean value is $\bar{x} = 24.25/17 = 1.426$ m. This is not quite as accurate as the previous answer because the values have been taken at the mid point of the band.

7. MEDIAN

Whenever you see words like, "the average person ...", or "the average income of ..." you don't always want to know the mean. Often you want to know the about the one in the middle. That's the **median**.

Again, this statistic is easy to determine because the median literally is the value in the middle. In order to find it, you just line up the values in your set of data from largest to smallest. The one in the dead-centre is your median. Our table would look like this.

Sample number	1	2	3	4	5	6	7	8	9	10	11	12	13
Height (m)	1.28	1.29	1.32	1.34	1.35	1.37	1.37	1.44	1.44	1.46	1.46	1.47	1.49

Sample number	14	15	16	17	Total
Height (m)	1.55	1.56	1.56	1.62	24.34

The mid point in the table is point number 9 so the median value is 1.44 m. If we had an even number of samples, say 18, then there would be two values in the middle and we should average the two to get the median.

8. MODE

The mode is the most frequently occurring value. In the example repeated below, this will be the class with a mark of 1.45 since there are six in this group.

x (mark)	1.25	1.35	1.45	1.55	1.65	
f.	2	5	6	3	1	total
f x	2.5	6.75	8.7	4.65	1.65	24.25

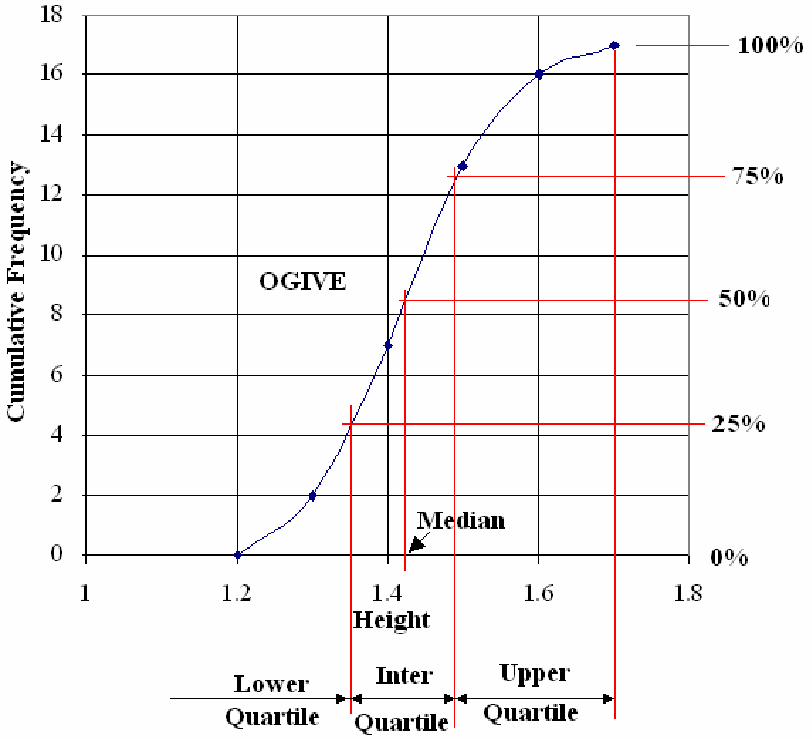
It is quite possible that the mode is not unique because the same maximum figure could occur more than once the distribution. If we don't have grouped data, there is no mode unless several occur at the same precise height. This leads us on to the next section.

9. OGIVE and QUARTILES

If we add a new row to our data showing the accumulative frequency and plot the data against it, we get a different sort of graph called an **OGIVE** that makes it easier to spot the median. Let's add a new row to our frequency distribution table containing the cumulative frequency. We can plot the same data using bands of 0.1 m. We plot the upper limit of each band so that each frequency shows how many children are either shorter or taller than that value. The table to plot is as follows.

The vertical scale is usually turned into % of the maximum as shown. The 50% level corresponds to the median.

x (upper point)	1.3	1.4	1.5	1.6	1.7	Total
f.	2	5	6	3	1	n =17
f x	2.5	6.75	8.7	4.65	1.65	24.25
cum. f	2	7	13	16	17	



The range that covers 75% to 100% is called the upper quartile. The range that covers 0% to 25% is called the lower quartile. The range between 25% and 75% is called the inter quartile and this can be divided into two parts called the semi-inter-quartile. These tell us something about how the samples are spread around the median but a better method of doing this is to use the standard deviation. Any range corresponding to a change of 1% is called a percentile.

WORKED EXAMPLE No.1

A company manufactures steel bars of nominal diameter 20 mm and cuts them into equal lengths. The diameter of each length is measured at the middle for the purpose of quality control. The results for 20 bars are given below.

- ***Produce a frequency distribution table using bands of 0.1 mm.***
- ***Calculate the mean of the samples.***
- ***Draw a histogram.***
- ***Plot the Ogive.***
- ***Determine the median, mode, upper and lower quartiles and the semi- inter-quartile.***

Use the tables and plots to show your solutions

Sample	1	2	3	4	5	6	7	8	9	10
diameter	19.9	19.8	20.1	19.9	19.7	20.1	20.0	19.6	19.7	20.1

Sample	11	12	13	14	15	16	17	18	19	20
diameter	20.2	20.0	19.9	19.8	20.1	20.0	19.7	19.6	19.9	20.2

SOLUTION

Total = 398.3 Total samples n = 20 mean = 398.3/20 =19.915

RANKED ORDER

Sample	1	2	3	4	5	6	7	8	9	10
diameter	19.6	19.6	19.7	19.7	19.7	19.8	19.8	19.9	19.9	19.9

Sample	11	12	13	14	15	16	17	18	19	20
diameter	19.9	20.0	20.0	20.0	20.1	20.1	20.1	20.1	20.2	20.2

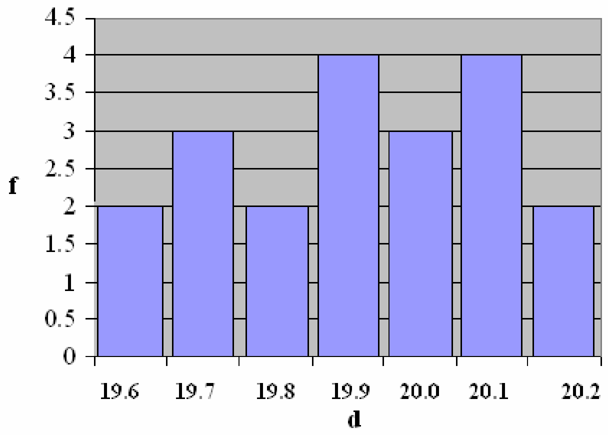
The median is the middle value of the samples when they are ranked in order and this is the 9th sample so the median is 19.9 mm.

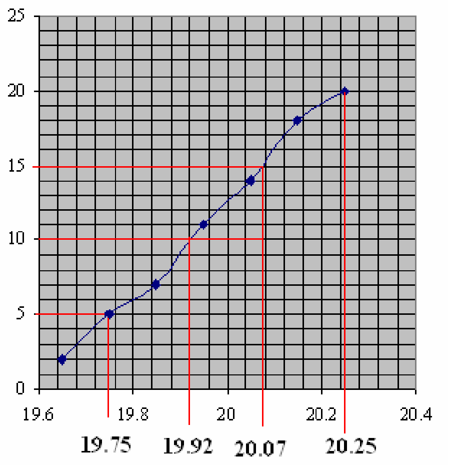
The most frequently occurring values are 19.9 and 20.1 so there is no unique mode.

Total = 398.3 Total samples n = 20 mean = 398.3/20=19.915

d	19.55-	19.65-	19.75-	19.85-	19.95-	20.05-	20.15- 20.25	Totals
Mark	19.6	19.7	19.8	19.9	20	20.1	20.2	
f	2	3	2	4	3	4	2	20
f d	39.2	59.1	39.6	79.6	60	80.4	40.4	398.3
cum.f	2	5	7	11	14	18	20	

n = 20 Mean = 398.3/20=19.915





Median =19.92

Upper quartile = 20.25 - 20.07 = 0.18   Lower quartile = 19.75 - 19.55 = 0.2

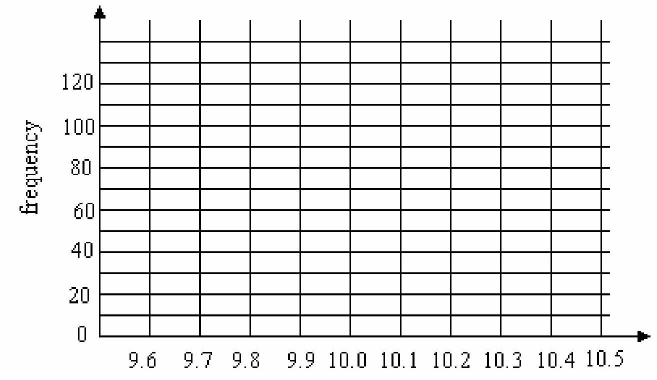
Inter-quartile range = 20.07 - 19.75 = 0.32   Semi-inter-quartile = 0.32/2 = 0.16

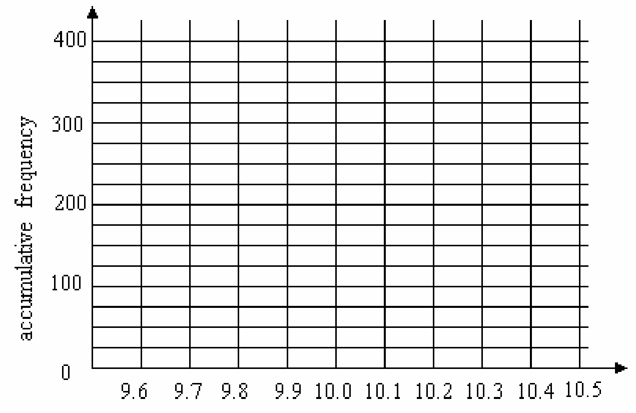
SELF ASSESSMENT EXERCISE No.1

The diameters of a number of components are measured to the nearest 0.1 mm. The distribution is shown.

Diameter mm	9.6	9.7	9.8	9.9	10.0	10.1	10.2	10.3	10.4
Number	3	9	36	88	122	90	44	7	1

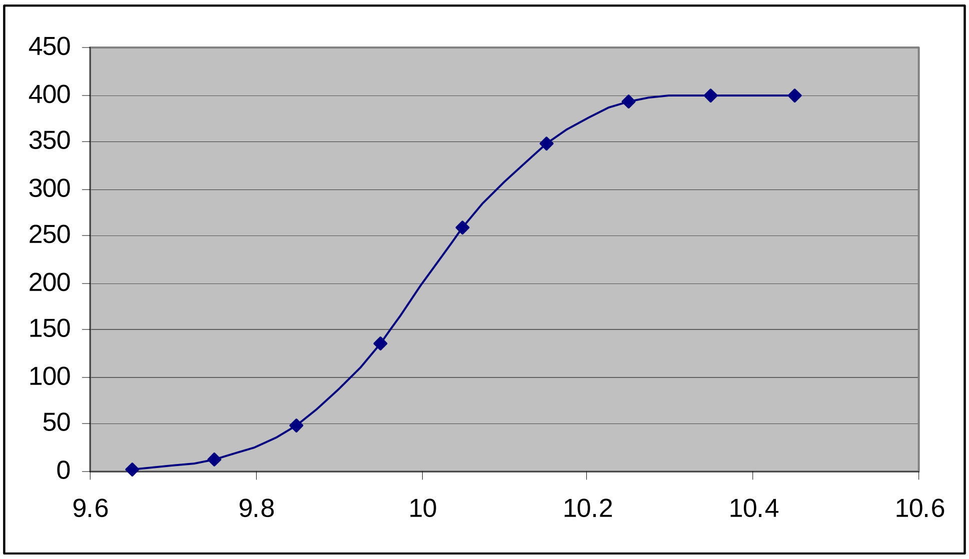
Draw the histogram and the Ogive and deduce the mean, the median, the upper and lower quartile and the semi-interquartile.





SOLUTIONS

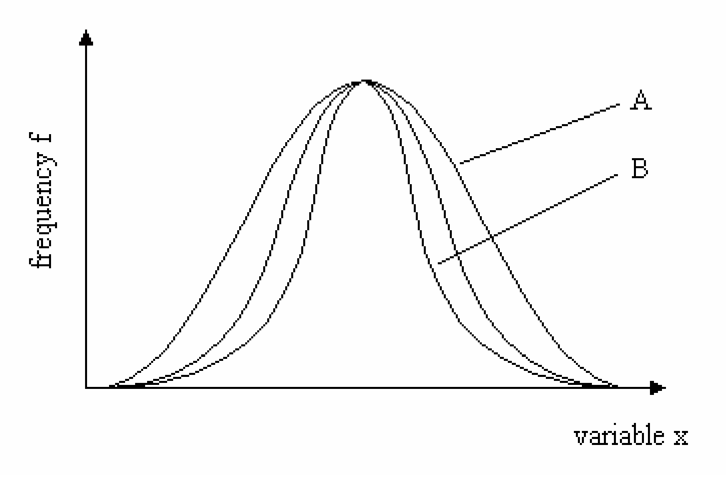
The ogive looks like this



(mean = 10.001 mm, median = 10 upper quartile = 3.6 lower quartile = 2.68 semi-interquartile = 0.9)

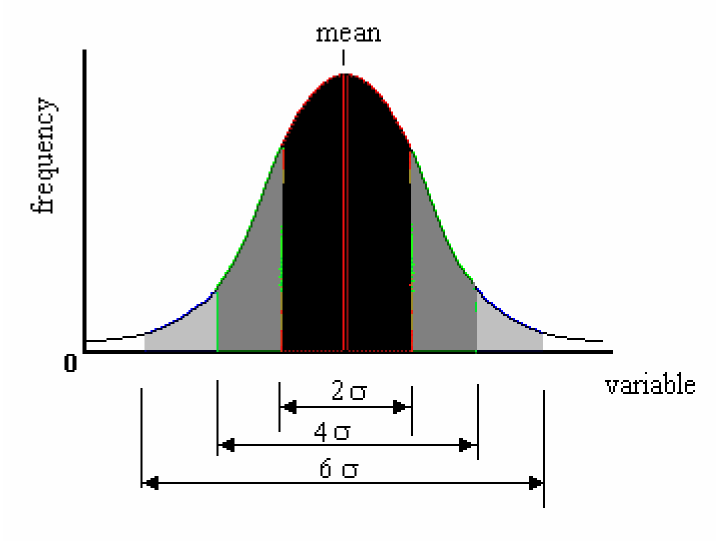
10. VARIANCE AND STANDARD DEVIATION

Standard deviation is a more difficult concept than the others we've covered. To understand this concept, it can help to learn about what statisticians call **normal distribution** of data. Most statistical samples produce a lot of results around the mean. Further away from the mean there are not so many results. The frequency distribution graph produces a "bell shaped" curve called the **normal distribution.** However, some are more normal than others and some times there are a lot more results close to the mean than others.



Consider the three cases shown on the diagram. The distribution shown by graph B has most of the examples in the set of data close to the "average," while those in graph A are more widely spread from the "average". In reality few examples tend to one extreme or the other. The **standard deviation "σ"** is a statistic that tells you how tightly all the various examples are clustered around the mean in a set of data.

When the examples are tightly bunched together and the bell-shaped curve is steep, the standard deviation is small (graph B). When the examples are spread apart and the bell curve is relatively flat, that tells you that you have a relatively large standard deviation (graph A). Let's first try to understand graphically what a standard deviation represents...



One standard deviation away from the mean in either direction on the horizontal axis (2σ) accounts for somewhere around 68 percent of the samples. Two standard deviations away from the mean (4σ), account for roughly 95 percent of the samples. Three standard deviations (6σ) account for about 99 percent of the samples. If this curve were flatter and more spread out, the standard deviation would have to be larger in order to account for the 68 percent so that's why the standard deviation can tell you 'how spread out' the examples in a set are from the mean.

This is useful in manufacturing as it tells us a lot about the quality of what you are making and how the equipment used in the process is behaving. So if for example you were monitoring the values of electrical resistors or the diameter of pistons being produced by machinery, a small standard deviation will tell us that they are being produced very accurately and close to the mean. Suppose all the samples between 4σ and 6σ are rejects. The larger the value of σ, the more rejects. Also if the mean of the sample is moving as time goes on, it means that more samples will be rejected on one side than the other and indicates that the something in the machine (e.g. the grinding wheel) is wearing away.

11. CALCULATION OF STANDARD DEVIATION for UNGROUPED DATA

Ungrouped data is presented in a table listing the value of each sample. If the number of samples is large, this becomes a large table but it is probably best to use this method with small numbers of samples.

Standard deviation σ = Square root of the mean of the variables squared.
The variance is denoted $S = \sigma^2$

$$S = \sigma^2 = \frac{\sum\left(x - \bar{x}\right)^2}{n-1} \quad n = \text{number of samples}$$

You might find it better to arrange your tables in columns rather than rows. Let's look at another example.

WORKED EXAMPLE No.2

The following is a table of lead concentration in the blood of a group of people. Calculate the mean and the standard deviation.

Sample	Resistance (Ohms)	Difference from mean	Differences squared
1	119	-1.8	3.24
2	120	-0.8	0.64
3	120	-0.8	0.64
4	121	+0.2	0.04
5	122	+1.2	1.44
6	119	-1.8	3.24
7	119	-1.8	3.24
8	122	+1.2	1.44
9	123	+2.2	4.84
10	123	+2.2	4.84
Totals10	1208		23.6

Mean = 1208/10 = 120.8

The sum of the squares of the differences (or deviations) from the mean, 23.6, is now divided by the total number of observation minus one, to give the ***variance.***

VARIANCE

$$S = \sigma^2 = \frac{\sum\left(x - \bar{x}\right)^2}{n-1} = \frac{23.6}{9} = 2.622$$

Finally, the square root of the variance provides the standard deviation:

$\sigma = 2.622^{½} = 1.619$ Ohms

SELF ASSESSMENT EXERCISE No.2

1. The hardness of ten steel samples was measured and the results were as follows.

Sample	1	2	3	4	5	6	7	8	9	10
Hardness	90	92	95	91	98	102	97	92	95	99

Calculate the mean and the standard deviation. Answer 95.1 and 3.9

2. The thickness of 20 steel strips was measured in mm and tabulated as shown.

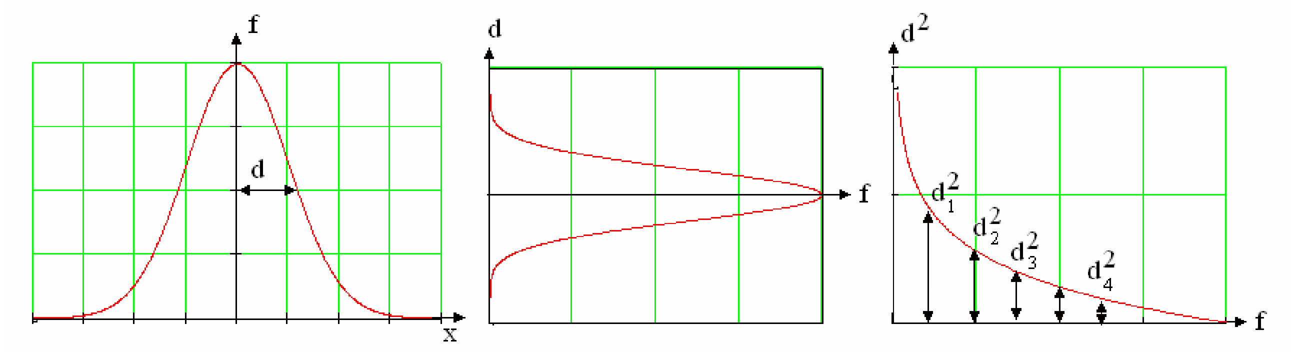
Sample	1	2	3	4	5	6	7	8	9	10
Thickness	19.8	19.9	19.9	20.1	20.1	19.9	20.2	19.7	19.7	19.9

Calculate the mean and the standard deviation. Answer 19.92 and 0.168

12. CALCULATION OF STANDARD DEVIATION for GROUPED DATA

Grouped data is presented in tables showing the bands and the frequency and is more likely to be used with large numbers of samples.

Consider a normal distribution curve. The mean occurs at the middle. The deviation from the mean at any point is d. Next consider the graph of d plotted against f and further the graph of d^2 plotted against f. On this last graph we find the mean d^2 as follows.



The area of the graph may be found from the mid-ordinates. $A = w(d_1^2 + d_2^2 + d_3^2 + ...d_n^2)$

The mean height of the graph is the variance $S = \dfrac{w(d_1^2 + d_2^2 + d_3^2 + ...d_n^2)}{f_n}$

w is the width of each strip = f_n/n and n is the number of strips. The mean height is then $S = \dfrac{f_n(d_1^2 + d_2^2 + d_3^2 + ...d_n^2)}{nf_n} = \dfrac{(d_1^2 + d_2^2 + d_3^2 + ...d_n^2)}{n}$

In general $S = \dfrac{\sum d^2}{n}$ For reasons not explained here, n-1 is often used instead of n on the bottom line. d is the deviation $d = x - \bar{x}$

$$S = \sigma^2 = \frac{\sum\left[f(x-\bar{x})^2\right]}{n-1}$$

σ is the standard deviation.

It can be shown that the formula simplifies to $\sigma^2 = \dfrac{\sum fx^2}{\sum f} - \left(\dfrac{\sum fx}{\sum f}\right)^2$

WORKED EXAMPLE No.3

The following is a grouped set of data for visits made to the doctor by a sample of children.

Visit to Doctor x	No.of Children f	Total Visits fx	Cumulative	x^2	fx^2
0	2	0	2	0	0
1	8	8	10	1	8
2	27	54	64	4	108
3	45	135	199	9	405
4	38	152	351	16	608
5	15	75	426	25	375
6	4	24	450	36	144
7	1	7	457	49	49

Totals $\Sigma f = n = 140$ $\Sigma f\,x = 455$ $\Sigma fx^2 = 1697$

Mean number of visits = 455/140 = 3.25.

$$\sigma^2 = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2 = \frac{1697}{140} - \left(\frac{455}{140}\right)^2 = 1.55$$

$$\sigma = 1.25$$

Some texts give the formula as $\sigma^2 = \dfrac{\sum fx^2}{\sum f - 1} - \left(\dfrac{\sum fx}{\sum f}\right)^2 = \dfrac{1697}{139} - \left(\dfrac{455}{140}\right)^2 = 1.57$

$$\sigma = 1.25$$

This does not make much difference so long as the total number of samples is very small.

WORKED EXAMPLE No.4

The hardness of 143 samples of steel is measured and grouped into bands as shown. Calculate the mean and standard deviation. The figures of 17.5 and 21.5 result from one sample being exactly 91 units and so half is allocated to each band.

Range	mid point	freq.	fx	acc f	x^2	$f\,x^2$
	x	f				
89-91	90	17.5	1575	17.5	8100	141750
91-93	92	21.5	1978	39	8464	181976
93-95	94	32	3008	71	8836	282752
95-97	96	38	3648	109	9216	350208
97-99	98	17	1666	126	9604	163268
99-101	100	17	1700	143	10000	170000
Totals		143	13575			1289954

Mean = 13575/143 = 94.93

$$\sigma^2 = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2 = \frac{1289954}{143} - \left(\frac{13575}{143}\right)^2 = 8.939$$

$\sigma = 2.99$

It is of interest to note that in this population, we get a very different answer using the other formula.

$$\sigma^2 = \frac{\sum fx^2}{\sum f - 1} - \left(\frac{\sum fx}{\sum f}\right)^2 = \frac{1289954}{142} - \left(\frac{13575}{143}\right)^2 = 72.46$$

$\sigma = 8.51$

SELF ASSESSMENT EXERCISE No.3

1. The accuracy of 100 instruments was measured as a percentage and the results were grouped into bands of 1% as shown. Calculate the mean and the standard deviation.

Range	Mid	freq
61.5-62.5	62	1
62.5-	63	2
63.5-	64	3
64.5-	65	4
65.5-	66	8
66.5-	67	12
67.5-	68	13
68.5-	69	18
69.5-	70	14
70.5-	71	10
71.5-	72	5
72.5-	73	4
73.5-	74	3
74.5-	75	2
75.5-76.5	76	1

Answers 68.88 and 2.74%

2. The breaking strengths of 150 spot welds was measured in Newton and grouped into bands of 20 N as shown.

Range	f
160-10	2
180-200	6
200-220	10
220-240	28
240-260	50
260-280	31
280-300	15
300-320	8

Calculate the mean and the standard deviation. (Answers 251.47 N and 29.04 N)