SUMMARY

TYPES OF QUANTITATIVE (NUMERICAL) DATA

- **Discrete data**, that takes exact number values, and is usually a result of counting.
  e.g., 1, 2, 3, 4, ...
  For example, the number of houses in a street.
- **Continuous data**, that takes number values within a certain range, and is usually a result of measurement.
  e.g., 5, 5.5, 5.55
  For example, the height of plants.

ORGANISING DATA

▶ **Ungrouped data**

- *Raw data:* 12, 21, 32, 15, 23, 34, 27
- *Ordered data:* 12, 15, 21, 23, 27, 32, 34
- Stem and leaf plot

1	2 5
2	1 3 7
3	2 4

Back to back stem and leaf plot

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

Note: 2 | 3 represents 23

- Frequency table

number	*frequency*
5	2
6	3
7	3
8	1

▶ **Grouped data**

- Frequency table

height (cm)	*freq*
0 - 4	5
5 - 9	9
10 - 14	15
15 - 19	6

Cumulative frequency table

height (cm)	*cumulative freq*
0 - 4	5
5 - 9	14
10 - 14	29
15 - 19	35

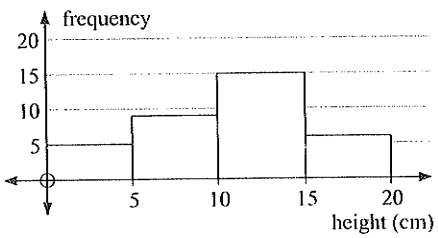
- Class intervals are equal width.
- Mid-interval values (class midpoints)
  $\frac{0+4}{2} = 2$, $\frac{5+9}{2} = 7$ etc.
- Lower and upper boundaries of the class interval:
  5 - 9 has boundaries 4.5 - 9.5

REPRESENTING DATA

All graphs require **labelled axes** and a **marked scale.**

- **Histograms**

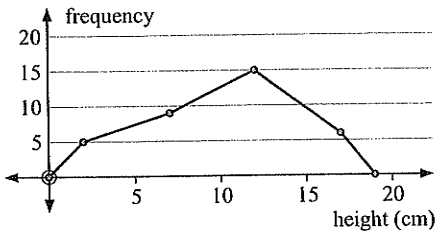
Height of plants



▶ Frequency is on the vertical axis.
▶ Bars are equal width.
▶ Bars are joined.

- **Frequency Polygons**

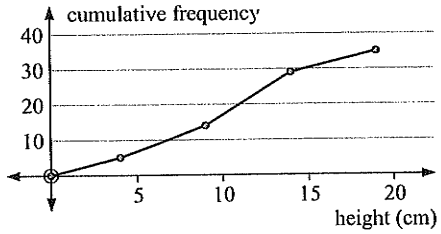
Height of plants



▶ Frequency is plotted against midpoint of class interval.
▶ Graph returns to the horizontal axes at the start of the first class and the end of the final class.

- **Cumulative frequency graphs**

Height of plants



▶ Cumulative frequencies plotted against the end points of the class interval.
▶ Curve begins at the horizontal axis.

MEASURES OF CENTRAL TENDENCY

Ungrouped discrete data: data such as 5, 6, 7, 2, 5

- The **mean**, $\overline{x}$, is the arithmetic average of the scores.
  For the given data, $\overline{x} = \frac{5+6+7+2+5}{5} = 5$
- The **median** is the middle value of an *ordered* set of data.
  The middle value is found at $\frac{n+1}{2}$
  where n is the number of terms in the data set.
  Since $n = 5$, $\frac{n+1}{2} = \frac{6}{2} = 3$ and as the ordered data set is: 2, 5, 5, 6, 7, the median = 5
- The **mode** is the most frequently occurring value
  $\therefore$ mode = 5.

Grouped discrete data:

This is data with a frequency table such as:

Number	*frequency*
5	2
6	3
7	3
8	1
Total	9

The mean is calculated by adding an fx column to the table.

Number (x)	*frequency (f)*	fx
5	2	10
6	3	18
7	3	21
8	1	8
Total	9	57

$$\text{mean} = \frac{\sum fx}{\sum f} = \frac{57}{9} \approx 6.33$$

As $\frac{n+1}{2} = \frac{9+1}{2} = 5$ the median is the 5th value
$\therefore$ median = 6.

The data is **bimodal** with modes of 6 and 7.

Grouped data

Frequency table

height (cm)	*midpoint (x)*	*freq (f)*	*fx*
0 - 4	2	5	10
5 - 9	7	9	63
10 - 14	12	15	180
15 - 19	17	6	102
	Total	35	355

The mean, $\bar{x} = \dfrac{\sum fx}{\sum f} = \dfrac{355}{35} \approx 10.1$

We use a **cumulative frequency polygon** to find the median.

The modal class is 10 - 14.

QUARTILES

Ungrouped discrete data e.g., 2, 5, 8, 9, 12, 15, 18, 19, 21

- As $n = 9$, the lower quartile Q_1 is found at

  $\dfrac{n+1}{4} = \dfrac{10}{4} = 2.5,$

  i.e., half way between the 2nd and 3rd values

  $\therefore \quad Q_1 = \dfrac{5+8}{2} = 6.5$

- The second quartile, Q_2, (median) is found at

  $\dfrac{n+1}{2} = \dfrac{10}{2} = 5,$

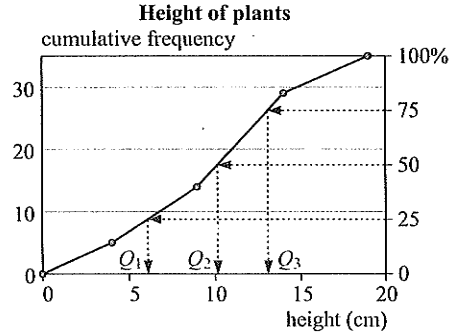
  i.e., the 5th score. $\therefore \quad Q_2 = 12$

- The upper quartile Q_3 is found at the $3 \times 2.5 = 7.5$ position.
  i.e. half way between the 7th and 8th values

  $\therefore \quad Q_3 = \dfrac{18+19}{2} = 18.5$

Grouped data

Quartiles (and **percentiles**) are found from a **Cumulative frequency graph**. (A percentile scale may be added to the graph.)



Notice that

- $Q_1 = 25$th percentile, $Q_2 = 50$th percentile, $Q_3 = 75$th percentile
- We draw lines on the graph to show the method.
- Percentile values can be easily read using the percentage scale.

MEASURES OF DISPERSION

For the **ungrouped discrete data**: 2, 5, 8, 9, 12, 15, 18, 19, 21

- **Range** = highest value − lowest value $= 21 - 2 = 19$
- **Interquartile range (IQR)** $= Q_3 - Q_1 = 18.5 - 6.5 = 12$
- **Standard deviation**

  We use graphics calculators to find the standard deviation in examinations. Always use the value for the population standard deviation in examinations (x_{σ_n} for Casio or σ_x for TI).

OUTLIERS

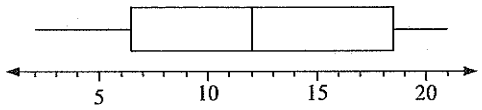
Outliers are values outside the limits of $Q_1 - 1.5 \times \text{IQR}$ and $Q_3 + 1.5 \times \text{IQR}$.

For our example,

$6.5 - 1.5 \times 12 = -11.5$ and $18.5 + 1.5 \times 12 = 36.5$, and as no data values lie outside the interval $-11.5 < x < 36.5$, there are no outliers in this data set.

DISPLAYING THE SPREAD OF DATA

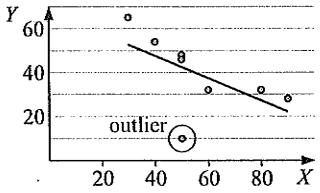
We use a **box and whisker plot**.



- A scale must be used.
- Label appropriately.
- The five-figure summary (min, Q_1, Q_2, Q_3, max) can be obtained from the box plot drawn on gdc (use trace).

BIVARIATE DATA (TWO VARIABLES)

We use a **scatter diagram** (or **scatter plot**).



Note that:

- Straight **line of best fit** (drawn by eye) must pass through $(\bar{x}, \bar{y})$.
- **Outliers** are often omitted from data and statistics recalculated.
- **Correlation** describes the nature and direction of any relationship.
- Correlation does not imply any causal relationship.

CORRELATION

To find the strength of the relationship between two variables we use **Pearson's product-moment correlation coefficient**, r.

- The calculated value (r) describes the strength of any relationship, $-1 \leqslant r \leqslant 1$.
- r close to ± 1 very strong relationship
  $r > 0.8$ or < -0.8 strong
  $r > 0.6$ or < -0.6 moderate
  r between -0.6 and 0.6 weak
- In the example above, the relationship would be weak/moderate ($r \approx 0.6$) with the outlier included, and strong ($r > 0.9$) with the outlier removed.
- In examination questions where a calculation for r is required, the value of the covariance s_{xy} will be given.
- Students are expected to find s_x and s_y using a gdc.
- If use of formula is not stipulated, students are expected to find the value for r using a gdc.

REGRESSION LINE FOR y ON x

You need to be able to find the regression line for y on x.

Note that:

- For calculation by formula, s_{xy} will be given in examination questions.
- If use of formula is not stipulated, students are expected to determine the equation of the regression line using a gcd.
- The regression equation is used for estimating values.
  It is usually reliable when **interpolating**.
  It can be unreliable when **extrapolating**.

CHI-SQUARED (χ^2) TEST FOR INDEPENDENCE

Notes:

- Write down null and alternate hypotheses.
- Calculation of χ^2 expected values by hand can be asked for.
- Calculation of the χ^2 statistic by formula is required.
- Use of tables χ^2 and the calculation of degrees of freedom.
- If calculation by formula and use of tables is not stipulated it is expected that students will determine values from a gdc.
- Knowledge and use of probability value (p-value) from a gdc.
- Examination questions will focus on the upper tail test.
- Expected values less than 5 result in an unreliable χ^2 test. We combine cells to overcome this problem.

TOPIC 6 – STATISTICS (SHORT QUESTIONS)

1 The number of customers entering a shop each hour on a particular day is listed below.

14, 23, 26, 34, 24, 18, 26, 16, 25

a Is the data discrete or continuous?

b Determine the mean, median, mode and the range for this data.

c Find the total income for the shop if the mean amount spent per customer is \$14.20

2 The number of houses in certain streets of a council area is presented in the following frequency table.

no. of houses	*frequency*
0 - 9	5
10 - 19	12
20 - 29	14
30 - 39	18
40 - 49	8
Total	57

a Is the data discrete or continuous?

b Construct the histogram for this data.

c On the same diagram, draw the frequency polygon for this data.

d What is the modal class for this data?

3 The stem plot below lists the number of birds present in a park on 22 days last month.

0	3, 5, 7
1	3, 4, 7, 7, 8
2	0, 2, 2, 2, 5, 7
3	1, 4, 5, 6, 9
4	0, 2, 9

scale: 1 | 3 represents 13 birds

a Determine the median and quartile values.

b Test for any outliers in this data.

c Find the probability that, on any day, the number of birds present is more than the upper quartile.

4 The height of 50 plants was measured (to the nearest cm) and the results are given in the table shown.

Height (cm)	*frequency*
0 - 9	2
10 - 19	15
20 - 29	21
30 - 39	7
40 - 49	5
Total	50

a Is the data discrete or continuous?

b List the mid-interval values for each class.

c Write down the lower and upper boundaries for the second class.

d Find the approximate mean height of the plants.

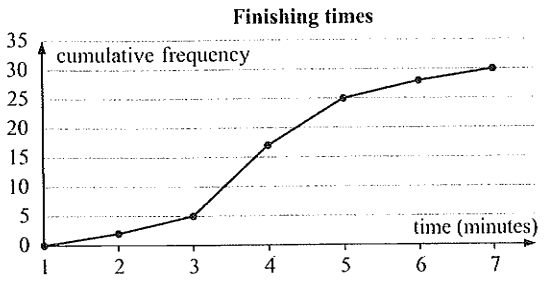
5 **a** Write down the cumulative frequencies for the height of plants in question **4**.

b Draw the cumulative frequency graph.

c Use your graph to determine the 80th percentile.

d How many plants are taller than the 80th percentile?

6 The Cumulative frequency curve below represents the finishing time (minutes) of 30 competitors in a recent orienteering contest.



Use the graph above to find:

a the median finishing time (Answer to 1 decimal place.)

b the time required for a runner to finish in the 1st Quartile.

c How many runners finished in a time between 2 and 3 minutes?

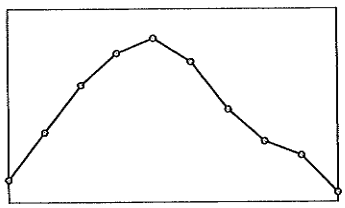
7 **a** The mean of 7 integers is 14. The integers, in ascending order, are: 9, 10, a, 13, b, 16, 21.
Find the values of a and b.

b Six integers, in ascending order, are: 1, 5, 9, 11, 16, p.
If the mean of the six numbers has the same value as the median, find p.

8 The statistics below represent a sample of 30 employees' wages (\$'000) at two firms.

Data set 1: mean = 38, median = 35, standard deviation = 7

Data set 2: mean = 38, median = 41, standard deviation = 11.5



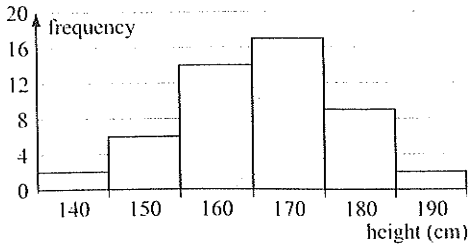
a The diagram above is a frequency polygon for data set 1. On the same diagram, sketch an approximate frequency polygon for data set 2.

b Which of the data sets has the greater dispersion in the wages paid to their employees?

c Which of the data sets is likely to have the smaller inter quartile range?

d Which of the data sets is likely to have more people earning higher wages?

9 A survey of the heights of Year 12 students at an international school gave the following results. All measurements have been rounded to the nearest 10 cm.



a What are the boundaries for the 170 cm class?

b Find the approximate mean and standard deviation of the students' height in this survey.

c How many students were taller than 2 standard deviations above the mean?

10 The list below shows the amount of weekly rent, in dollars, for houses in a certain city.

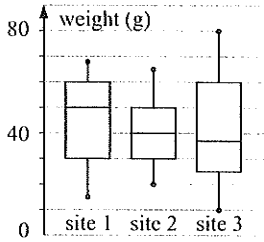
Weekly rent ($)	*Frequency*
80 - 99	3
100 - 119	15
120 - 139	26
140 - 159	30
160 - 179	14
180 - 199	1

a Find the mean and standard deviation for weekly rents.

b What is the probability that the rent for a randomly chosen house will be greater than $140?

c Determine the percentage of houses that have rents greater than one standard deviation above the mean.

11 The box plots show the weights of particular species of fungus collected from 3 different sites in a forest.



a Which site has the greatest range of weights?

b At which site are the weights of fungi least spread?

c Which site has the highest median weight?

d At which site were the heaviest fungi found?

e Which site has the highest proportion of weights above 40 grams?

f Which site has the lowest proportion of weights above the upper quartile?

12 For the boxplot shown below:

a write down the values of the lower quartile, median and upper quartile

b find the range

c calculate the value of the interquartile range

d determine whether the minimum value is an outlier.

13 The following marks were obtained by students in a Mathematics Examination.

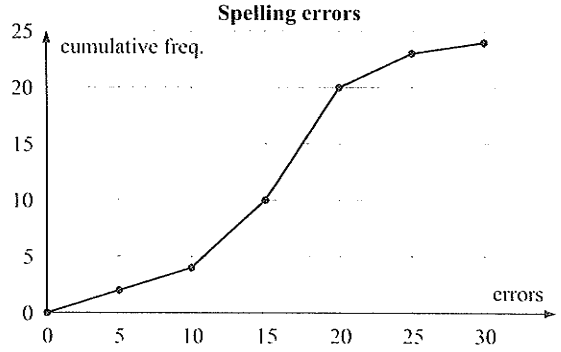
67 62 75 78 78 49 57 59 61 72
75 25 82 68 85 81 48 70 76 87

a Find the five-number summary.

b Represent this information as a box and whisker plot.

c Test for any outliers. Suggest a possible reason for any outlying value.

14 The numbers of spelling errors made by year 9 students in an essay are displayed on the cumulative frequency diagram.



a Use the cumulative frequency graph to find the five-number summary for the numbers of errors students made in the essay.

b Draw the box and whisker plot for student spelling errors.

15 The following prices ($'000) for houses sold last month in a certain suburb were:

240, 260, 262, 280, 310, 325, 330, 340, 760

a Find the median price for these sales.

b Determine the interquartile range.

c Omit any outliers from the data and recalculate the median.

d What is the percentage change in the median price with the outlier omitted?

16 Given: $s_x = 17.4$, $s_y = 25.6$, $s_{xy} = 405$, $\overline{x} = 63$, $\overline{y} = 110$:

a Calculate the coefficient of correlation (r).

b Describe the relationship between the two variables, x and y.

c Determine the equation of linear regression for y on x.

d Find the value of y when x is 70.

17 The table shows the exchange rate for Argentine pesos (against USD) and interest rates in Argentina over a period of time.

Exchange rate	2.85	2.95	2.90	2.75	2.65
Interest rate	7.40	7.50	7.55	7.25	7.25
Exchange rate	2.80	3.05	2.98	2.95	
Interest rate	7.35	7.65	7.75	7.60	

a Draw the scatter diagram for the given data.

b Write down the value of the correlation coefficient for this data.

c Describe the nature and strength of the relationship that appears to exist between the exchange rate and interest rates over this period of time.

18 Consider the following data on farm production.

Monthly rainfall (mm)	5	10	15	20	25	30
Yield (tonnes)	14	21	29	31	30	28

a Determine the coefficient of correlation (r). What does the value of r suggest about the nature and strength of the relationship between monthly rainfall and crop yield?

b Draw the scatter diagram for the farm data.

c Explain why the coefficient of correlation may not be an appropriate measure for this data.

19 Consider the following contingency table.

	Y_1	Y_2
X_1	32	14
X_2	25	19

a Show that the expected values (whole numbers) are:

	Y_1	Y_2
X_1	29	17
X_2	28	16

b Show that the chi-squared statistic for this data is 1.72

20 A nursery has developed a new hybrid plant. They claim that this particular hybrid will grow equally well under any light conditions. They have provided the following data to support their belief.

	Height < 60 *cm*	*Height* $\geqslant 60$ *cm*
Sunlight	37	43
Shade	22	18
Dark	25	19

a Write suitable null and alternate hypotheses for a chi-square test.

b Write down the χ^2 statistic for the plant data.

c Write down the critical value at the 5% level of significance.

d Is there evidence to support the nursery's claim?

21 **a** Test the independence of the following factors at the 5% level of significance.

	Factor Y_a	Factor Y_b	Factor Y_c
Factor X_a	6	3	7
Factor X_b	21	35	28
Factor X_c	16	11	22

If expected values are less than 5, the reliability of the chi-square test is reduced.

b Identify the row/column containing the low expected value.

c Combine the first two rows and retest for independence.

d Comment on the result.

22 The number and weight of potatoes in a sample of 2.5 kg bags is listed.

Number in bag	89	97	105	110
Median weight (g)	28	26	25	23
Number in bag	125	140	145	150
Median weight (g)	21	18	18	16

a Determine the equation of linear regression for this information.

b Use your equation to find the number of potatoes in a bag if the median weight is:

i 100 grams **ii** 200 grams

c Which of the answers in **b** is likely to be more reliable? Give a reason for your answer.

23

X	30	50	80	60	50	90	40	50
Y	65	12	28	42	46	26	54	48

a Write down the equation of linear regression and the coefficient of correlation.

b Remove the outlier from the data and recalculate the equation of the regression line and the correlation coefficient.

c Comment on the change in the slope of the line and the strength of the relationship.

24 9 students sat a mathematics examination. The results that they obtained and the number of hours that each of them studied are shown in the table.

Study time (hrs)	7	6	3	16	15	11	18	32	20
Result (%)	56	42	25	80	65	60	85	96	90

a Write down the equation of the straight line of best fit.

b Tony's score in the examination was 70%. According to the result of the line of best fit, for how long did he study?

c In terms of the marks obtained in the examination, explain the meaning of the y- intercept and the gradient of the equation of the line of best fit.

25 A government agency believes that the evidence submitted by a chemical firm in support of the claim that their product is safe has possibly been manipulated. The agency asks you to conduct a chi-squared test.

	Minimal effect	*Negligible effect*	*no effect*
Plants without disease	101	109	115
Plants with disease	205	221	229

H_0: the deviation from expected values is due to random chance.

a Determine the chi-squared probability value for this data.

b Test the p-value against the 1% and 0.5% levels of significance (lower tail test).

c What conclusion can you draw from the chemical company's data?

TOPIC 6 – STATISTICS (LONG QUESTIONS)

1 The prices ($) for 25 similar printers are displayed in a stem and leaf diagram below:

27	1 2 4 9
28	0 1 5 7 8
29	0 0 2 3 4 7 8 9 9 9
30	3 4 7 9
31	1 6

Scale 28 | 5 represents $285

The mean of this data is 293 and the standard deviation 12.0

a **i** Calculate the median, range, lower and upper quartiles.

ii Display these statistics on a horizontal box and whisker plot. Use a scale of 1 cm to represent $10.

b Three months later the prices for the same printers were recorded:

The prices ranged between 269 and 329 dollars with a mean of 295 dollars and standard deviation of 11 dollars. The lower and upper quartiles were 280 and 305 respectively and the median was 295. Show this data as a box and whisker plot using the same scale as in **a**.

c **i** Describe the main difference between the box and whisker plots.

ii Explain whether or not this information shows that the price of printers has increased. Give a clear reason for your answer.

d Find the percentage increase in the mean price of printers over the 3 month period.

2 The following data was obtained in a statistical experiment which involved measuring the distance travelled by two toy cars. Each car was rolled down a slope 40 times. The measurements were rounded to the nearest tenth of a metre.

Red car

3.6	4.6	5.6	6.4	4.2	5.3	6.1	4.5
5.4	4.6	3.9	6.2	5.8	4.5	5.4	6.1
4.5	5.6	5.7	4.8	3.9	5.6	6.1	5.9
4.1	5.3	4.2	6.2	7.4	5.4	5.8	4.5
3.9	5.4	5.7	4.8	5.4	5.7	6.1	6.4

Blue car

number of rolls	40
Mean distance	4.9 m
Median distance	4.8 m
Shortest distance	3.2 m
Longest distance	6.7 m
First quartile	4.1 m
Third quartile	5.4 m
Standard deviation	0.8 m

a Determine the mean and standard deviation for the distance travelled by the red car.

b Complete the table of cumulative frequencies for the red car data.

Distance (m)	*Cumulative frequency*
3.5 - < 4.0	
4.0 - < 4.5	
4.5 - < 5.0	
5.0 - < 5.5	
5.5 - < 6.0	
6.0 - < 6.5	
6.5 - < 7.0	
7.0 - < 7.5	

c Draw the Cumulative frequency graph for the distance travelled by the **red car**. Use a scale of 1 cm to represent 1 m on the horizontal axis and 1 cm to represent 10 units on the vertical scale.

d Use the graph to find the following statistics for the **red car**:
- **i** median distance
- **ii** lower quartile
- **iii** upper quartile.

e Draw the box and whisker plots for both cars *on the same axis*.

f Compare the statistics for distance travelled by the two toy cars. Is it reasonable to assume that the same machine manufactured these two toys? Give reasons for your answer.

3 A manufacturer states that each box of a certain cereal contains 320 g, on average. Each box of a random sample of 24 boxes was weighed with the following results recorded, in grams.

312	320	326	330	306	322	326	330
312	308	307	316	315	328	334	309
308	325	320	332	316	321	314	324

a Calculate the mean weight and the range of weights for the boxes.

b Organise the data into a frequency table with the first class as 305 - 309.

c Use the information in your table to draw the frequency polygon for the cereal data.

d Comment on the manufacturer's stated average weight.
Six months later, another randomly selected 24 boxes were weighed with the following results.

Ave. weight	*Frequency*
310 - 314	3
315 - 319	5
320 - 324	8
325 - 329	6
330 - 334	2

e Calculate the mean weight for the new data.

f Draw the frequency polygon for the new data on the same graph as above.

g Does the evidence suggest that the manufacturer has improved the production process in the six months?

4 A large store employs 100 sales staff. The employees' total sales for last year are listed in the table below.

Sales ($'000)	*Number of Staff*
60 to less than 70	3
70 to less than 80	8
80 to less than 90	11
90 to less than 100	23
100 to less than 110	27
110 to less than 120	20
120 to less than 130	8

a Represent this information as a histogram.

b Write down the minimum and maximum sales required for a staff member to be in the highest class.

c Calculate the mean and the standard deviation of the sales per staff for the year using 65, 75, etc., as the midpoints of each class. Give your answers to the nearest hundred dollars.

d The store had an incentive scheme last year that offered a $900 bonus to all staff with sales exceeding +2 standard deviations from the mean.
- **i** What was the minimum amount of sales (S) required for a staff member to qualify for this bonus?
- **ii** Approximately how much money did the store pay in bonuses?

e If the top 8 sales staff were to get the bonus, how many standard deviations above the mean would the limit need to be set?

f At the end of the year the store's manager decided to reduce the number of sales staff. Every staff member whose sales were less than 1.385 standard deviations from the mean would be removed from the team. How many staff would go?

5 A jeweller measured the volume and mass of some samples of silver which he had purchased. He suspected that one of the samples might be a fake. The results are listed in the table.

Sample	A	B	C	D	E	F
Volume (cm^3)	3	6	4	7	16	8
Mass (g)	40	95	50	160	285	130
Sample	G	H	I	J	K	L
Volume (cm^3)	5	12	9	6	10	11
Mass (g)	65	210	155	90	170	190

a Draw the scatter plot for this data. Use 1 cm to represent 20 g on the horizontal axis and 1 cm to represent 1 cm^3 on the vertical scale.

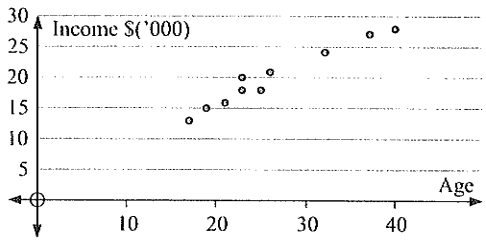
b **i** Calculate the mean for both volume ($\overline{x}$) and mass ($\overline{y}$).
ii Draw the straight line of best fit for the data. The line should pass through the point ($\overline{x}$, $\overline{y}$).

c Describe the relationship which appears to exist between the volume and mass of the samples of silver.

d **i** Write down the value of the (linear) coefficient of correlation.

e **i** Remove the suspect value from the data and then write down the equation of the linear regression line for this data.
ii Use your equation to find the expected mass of the sample of silver with the same volume as the suspect sample.
iii Calculate the percentage error between the given and expected masses of the suspect sample based on the expected mass.

6 The scatter diagram shows the age and annual income for 10 randomly chosen individuals.



The following statistics have been calculated for this data.

Age		*Income*	
Mean	26.3	Mean	20
Median	24	Median	19
Mode	23	Mode	18
Standard Deviation	7.25	Standard Deviation	4.78

The covariance for age and income is 33.9.

a **i** Determine the value of the correlation coefficient (r).
ii Describe the relationship between the age and annual income for these individuals.

b Determine the equation of the linear regression line for age and income.

c Use the equation in **b** to estimate:
- **i** the annual income for someone aged 30 years
- **ii** the annual income for someone who is 60 years of age.

d Comment on the reliability of both answers in **c**.

7 Fifty words were randomly selected from story A and another fifty words were randomly selected from story B. The number of letters from each word were recorded and the results are summarized in the frequency table below.

Number of letters in a word	*Frequency of words in story A*	*Frequency of words in story B*
2	5	7
3	13	9
4	6	10
5	8	6
6	5	6
7	5	8
8	2	2
9	4	1
10	2	1

a Draw frequency polygons for both stories on the same axes.

b Determine the five-number summary of word length for both stories and display the statistics as side-by-side box and whisker plots.

c Write down the mean and standard deviation of word length for both stories.

d Construct a 2×2 contingency table by adding the number of words with less than 5 letters and the number of words with 5 or more letters for each story.

e Test for independence at the 10% significance level for the chi-square distribution.

f What conclusion can be drawn from the test?

g Is there any evidence to support the claim that the two stories were written by different authors?

8 Jack believes that students' choice of breakfast cereal is related to gender. He conducts a survey at his school and records the following information. He plans to use a chi-squared test to discover if his belief is correct.

	Muesli	*Rolled Oats*	*Corn Flakes*	*Weetbix*
Female	2	7	24	12
Male	4	15	13	13

a Write a suitable null hypothesis for a chi-squared test.

b Write down the expected values for this data.

c Jack observes the expected values and realises that the results of a chi-square test may not be reliable because there are expected values which are less than 5. He decides to combine the first two columns of data in the contingency table.

Write down the contingency table for the combined data.

d Show that the chi-squared statistic for the combined table is 6.88.

e Test the calculated chi-squared statistic at the 5% level of significance.

f What conclusion can Jack draw from the test?

g If Jack had chosen to use the original contingency table, would he have drawn the same conclusion? Justify your answer.

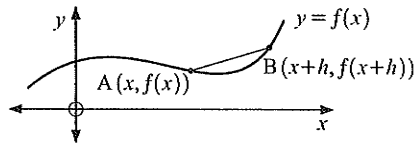
SUMMARY

NOTATION

- If $f(x)$ is a function, then $f'(x)$ is its derivative.
- If y is the graph of a function, then $\frac{dy}{dx}$ is its gradient function.

APPROXIMATING THE GRADIENT OF A TANGENT TO $y = f(x)$

For the graph of a function $y = f(x)$, consider two points A and B on $y = f(x)$, such that their coordinates are $A(x, f(x))$ and $B(x+h, f(x+h))$.



The gradient between A and B is $\frac{f(x+h)-f(x)}{h}$.

The gradient between A and B can be used as an approximation to the gradient of the tangent to $y = f(x)$ at the point A. As the point A is brought closer to the point B, the approximation improves.

FUNCTIONS OF THE FORM $y = ax^n$

For functions of the form $f(x) = ax^n$, where a is a constant, the **derivative** or **gradient function** is

$$f'(x) = anx^{n-1}.$$

Example:

Given $f(x) = x^3 - \frac{4}{x} + 5$, find $f'(x)$.

Solution:

$$f(x) = x^3 - \frac{4}{x} + 5$$
$$\therefore \quad f(x) = x^3 - 4x^{-1} + 5$$
$$\therefore \quad f'(x) = 3x^2 - 4(-1)x^{-2} + 0$$
$$\therefore \quad f'(x) = 3x^2 + \frac{4}{x^2}$$

GRADIENT AT A POINT

The gradient of $y = f(x)$ at a point $x = k$ is found by evaluating the derivative of $f(x)$ at $x = k$, i.e., by finding $f'(k)$.

Example:

Given $f(x) = 3x^2 - 2x + 1$, find the gradient of the function at the point where $x = 3$.

Solution:

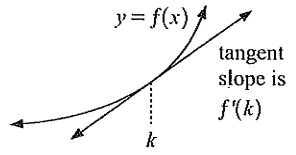
Differentiating $f(x)$, we have $f'(x) = 6x - 2$.

At $x = 3$, $f'(3) = 6(3) - 2 = 16$.

Hence the gradient at $x = 3$ is 16.

TANGENTS TO CURVES

To find the equation of the tangent to the curve $y = f(x)$ at $x = k$ we:



- Find the coordinates of the *point of contact*.

  At $x = k$, the point of contact is $(k, f(k))$.

- Find the *gradient function* $\frac{dy}{dx}$ (or $f'(x)$) and evaluate the gradient function at $x = k$.
- Determine the *equation of the line* using the point of contact $(k, f(k))$ and the gradient found at $x = k$.