Big Book of Helpful Hints for Yr 12 General Maths

1 : INVESTIGATING DATA DISTRIBUTIONS

1A Types of data

Categorical		Numerical	
Data fits into categories.		Data has been counted or measured	
Nominal	Ordinal	Discrete	Continuous
Yellow, Red, Blue	Small, Medium Large Star rating ★★★ ★★ ★	Number of siblings Days off sick	Height

1B Display and describe categorical variables

Frequency tables

list the values in a data set, and how often each value occurs.

- frequency: the number of times a value occurs
- percentage frequency : the percentage of times a value occurs, = $\frac{\text{count}}{\text{total}} \times 100$

Example : thirty children chose a sandwich, a salad or a pie for lunch, as follows:

sandwich × 7

salad × 10

pie × 13

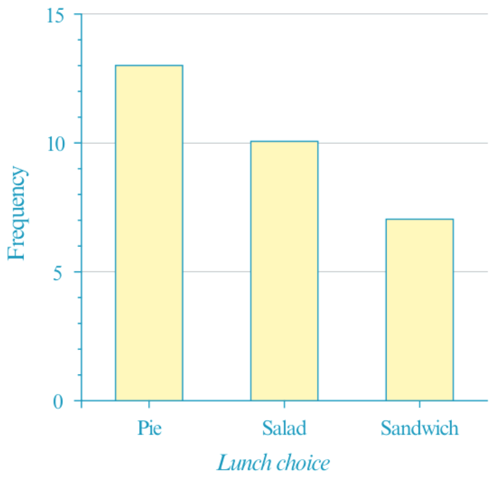
Lunch choice	Frequency	
	Number	%
Sandwich	7	23.3
Salad	10	33.3
Pie	13	43.3
Total	30	99.9 ≈ 100

Bar charts

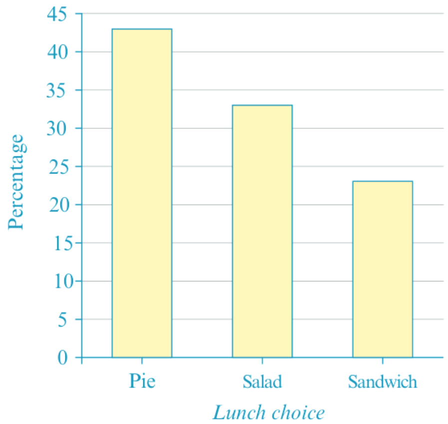
are for categorical data

- frequency or percentage frequency is shown on the vertical axis
- the variable being displayed is plotted on the horizontal axis
- the height of the bar (column) gives the frequency (or percentage)
- the bars are drawn with gaps to indicate that each value is a separate category
- there is one bar for each category.

Bar chart

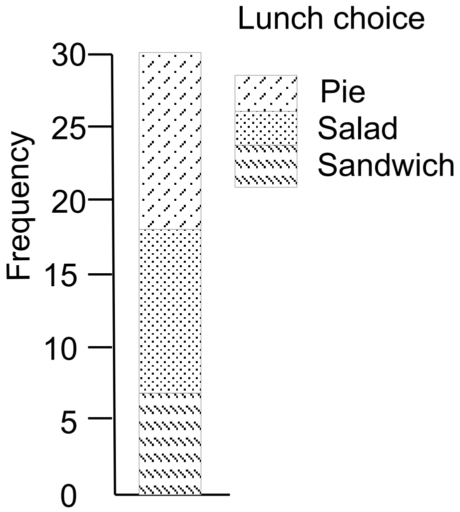


Percentage bar chart

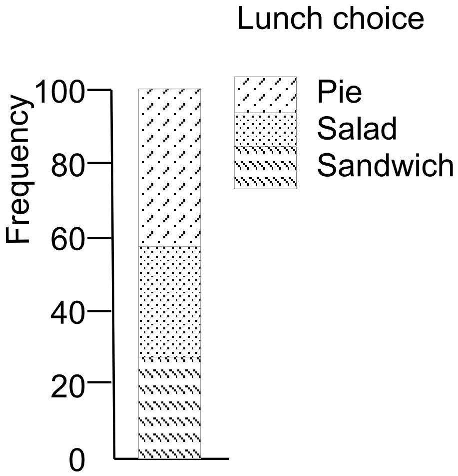


Segmented bar chart

Segmented bar chart



Percentage segmented bar chart



Mode, or modal category

is the most frequently occurring category.

Categorical variable report guidelines

■ Briefly summarise the context in which the data were collected including the number of individuals involved in the study.

■ If there is a clear modal category, ensure that it is mentioned.

■ Include frequencies or percentages in the report. Percentages are preferred.

■ If there are a lot of categories, it is not necessary to mention every category, but the modal category should always be mentioned.

1C Displaying and describing numerical data

Frequency tables

Discrete data, small number of values

use each value as a category

30 Year 11 students report the following number of siblings :

2 3 4 0 3 2 3 0 4 1 0 0 1 2 3 0 2 1 1 4 5 3 2 5 6 1 1 1 0 2

% frequency = $\frac{\text{Number}}{\text{Total}}$ e.g $\frac{6}{30} \times 100 = 20.0$

Number of siblings	Frequency	
	Number	%
0	6	20.0
1	7	23.3
2	6	20.0
3	5	16.7
4	3	10.0
5	2	6.7
6	1	3.3
Total	30	100.0

Grouped frequency table

when data has a large range of values or when the variable is continuous

The data below give the average hours worked per week in 23 countries.

35.0 48.0 45.0 43.0 38.2 50.0 39.8 40.7

40.0 50.0 35.4 38.8 40.2 45.0 45.0 40.0

43.0 48.8 43.3 53.1 35.6 44.1 34.8

% frequency = $\frac{\text{Number}}{\text{Total}}$ e.g $\frac{1}{23} \times 100 = 4.3\%$

Average Frequency hours worked	Frequency	
	Number	%
30.0 – 34.9	1	4.3
35.0 – 39.9	6	26.1
40.0 – 44.9	8	34.8
45.0 – 49.9	5	21.7
50.0 – 54.9	3	13.0
Total	23	99.9

Histograms

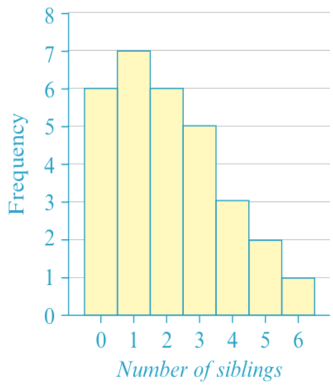
are for numerical variables.

■ frequency (number or percentage) is shown on the vertical axis

■ each column corresponds to a data value, or data interval

BUT ungrouped discrete data has the actual data value located at the middle of the column

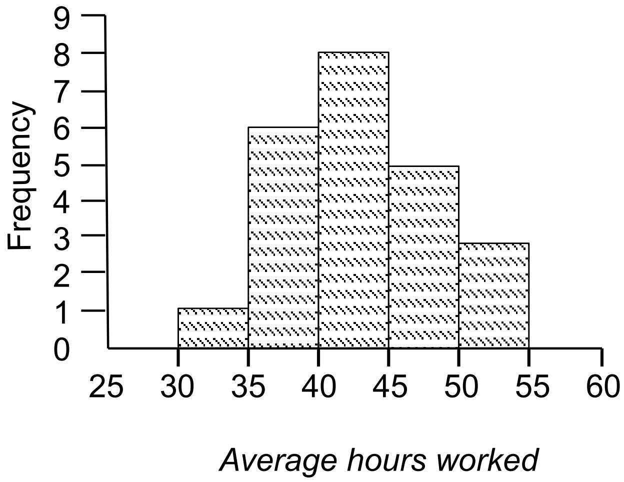
Histogram for ungrouped discrete data



actual value at middle of each column

Number of Siblings	Frequency
0	6
1	7
2	6
3	5
4	3
5	2
6	1
Total	30

Histogram from frequency table

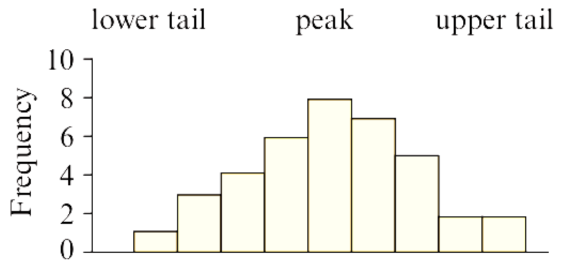


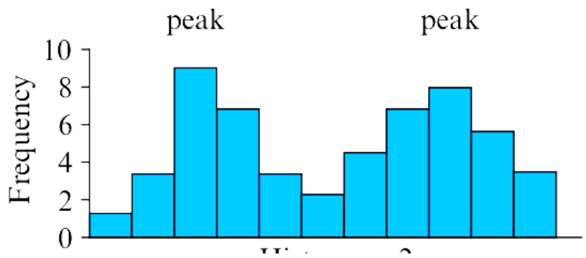
Average Frequency hours worked	Frequency	
	Number	%
30.0 – 34.9	1	4.3
35.0 – 39.9	6	26.1
40.0 – 44.9	8	34.8
45.0 – 49.9	5	21.7
50.0 – 54.9	3	13.0
Total	23	99.9

Report histogram

Shape

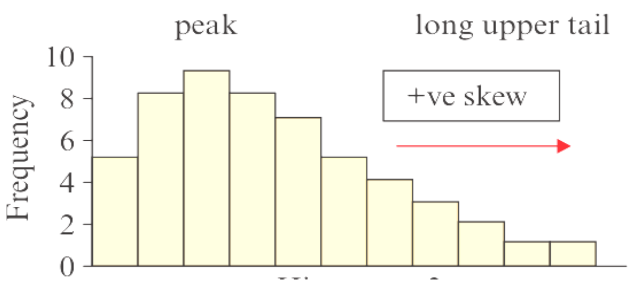
Symmetric distributions

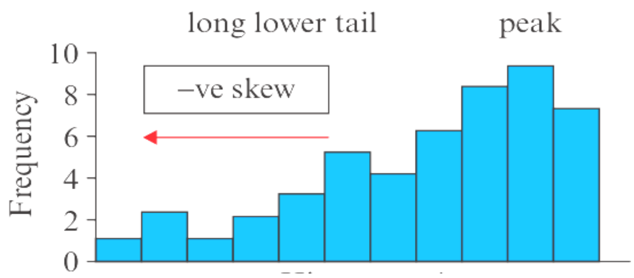




Single peaked symmetric distribution.	Double peaked symmetric distribution.
intelligence test scores weights of oranges any other data for which the values vary evenly around some central value.	Bimodal distribution. data from two different populations : distance thrown by male and female discus throwers.

Skewed distributions

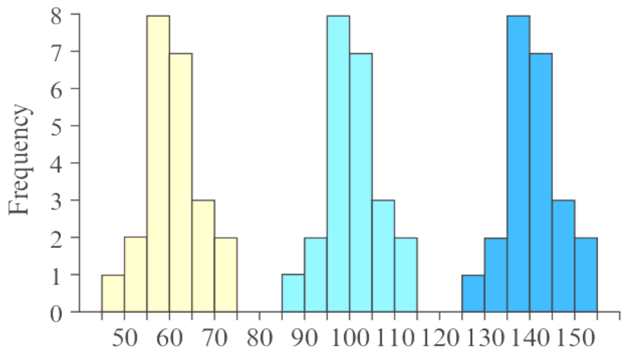




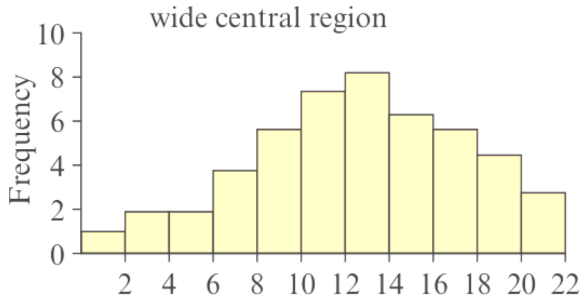
Positive skew	Negative skew
house prices pay rates in a large company	age at death

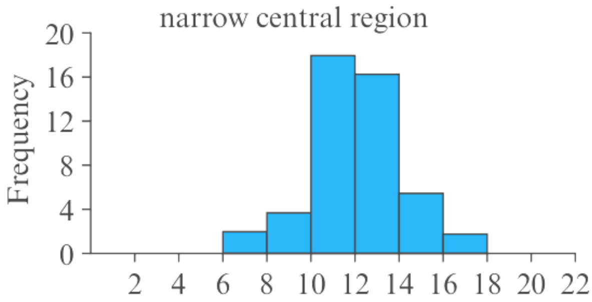
Centre

Identical in shape, these distributions are 'centred' at different points along the axis.

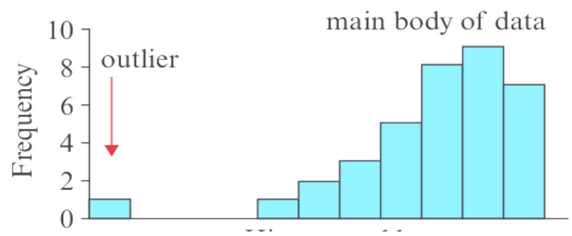


Spread





Outliers



1D Dot plots and stem plots

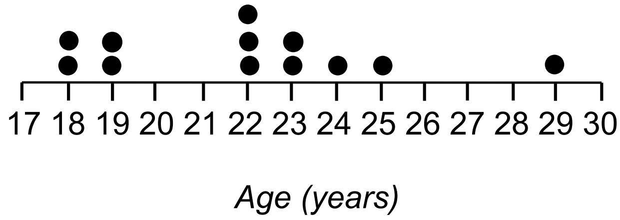
Dot plots

Dot plots display fairly small data sets where the data takes a limited number of values.

The ages (in years) of a cricket team are:

22 19 18 19 23 25 22 29 18 22 23 24 22

Construct a dot plot of these data.



Stem plots

Test results

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

Key : 1 | 5 = 15

The last digit is the leaf.

Stem plots with split stems

Generally the stem is split into halves or fifths as shown.

Marks obtained by 17 VCE students on a statistics test.

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

Key : 1 | 5 = 15

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

Key : 1 | 5 = 15

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

Key : 1 | 5 = 15

1E Using a $\log_{(10)}$ scale to display data

$\log_{10}x$

If $\log_{10}x = b$ then $10^b = x$

$\log_{10}(100) = 2$, because $10^2 = 100$

$\log_{10}(1000) = 3$, because $10^3 = 1000$

Logarithmic transformation

involves changing the scale on the horizontal axis from x to $\log_{10}(x)$, and replacing each of the data values with its logarithm.

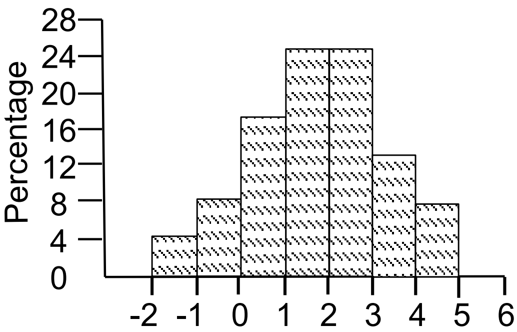
Working with logarithms

Find the log of 45. $\log_{10}(45) = 1.65...$

Find the number with log equal to 2.7125. $10^{2.7125} = 515.82...$

Analysing data displays with a logarithmic scale

Histogram of the weights of 27 animal species plotted on a log scale.



a) What body weight is represented by the number 4 on the log scale?

$10^4 = 10,000$

b) How many of these animals have body weights more than 10 000kg?

Last column on the right. 7.4% of 27 = 2.

c) The weight of a cat is 3.3 kg. Use your calculator to determine the log of its weight.

$\log_{10}(3.3) = 0.518...$

d) Determine the weight (in kg) of the animal with a log(body weight) of 3.4 (the elephant).

$10^{3.4} = 2511.88...$

1F Measures of centre and spread

Median, range and interquartile range

Median

is the middle value in an ordered data set.

For n data values the median is located at the $\left(\frac{n+1}{2}\right)$ data position,

odd number of data   2 3 4 5 5 5 6 7 8 8 11

Median value at $\frac{11+1}{2}$ = 6^{th} position.   Median = 5

even number of data   2 3 4 5 5 6 7 7 8 8 11 11

Median value average of values at $\frac{n}{2}$ and $\frac{n}{2}+1$ = 6^{th} and 7^{th} positions. Median = $\frac{6+7}{2}$ = 6.5

Range

R, is the difference between the largest and smallest values in the data set.

R = largest data value − smallest data value

Probs with range :

Because the range depends only on the two extreme values in the data, it is not always an informative measure of spread. For example, one or other of these two values might be an outlier.
And, any data with the same highest and lowest values will have the same range, irrespective of the way in which the data are spread out in between.

Interquartile range (IQR)

To measure the spread of a data distribution around the median (M) we use the interquartile range (IQR).

■ arrange all observations in order according to size

■ divide the observations into two equal-sized groups, and if n is odd, omit the median from both groups

■ locate Q1, the first quartile, which is the median of the lower half of the observations

■ locate Q3, the third quartile, which is the median of the upper half of the observations.

The interquartile range IQR is then: IQR = Q3 − Q1

Why is IQR more useful than range?

The IQR is a measure of spread of a distribution based on the middle 50% of observations. Since the upper 25% and lower 25% of observations are discarded, the interquartile range is generally not affected by the presence of outliers.

The mean and standard deviation

Mean $\bar{x}$

means average

$\bar{x}$ means add up all the (Σ) numbers (x) and then divide by how many there are (n)

$$\bar{x} = \frac{\Sigma x}{n}$$

mean of 4, 5 and 6 is $\frac{4 + 5 + 6}{3} = 5$

Choosing between the mean and the median

■ symmetric and no outliers : mean or median

■ clearly skewed and/or there are outliers : median

Standard deviation (s)

measures the spread of a distribution about the mean ($\bar{x}$)

$$s = \sqrt{\left(\frac{\Sigma(x-\bar{x})^2}{n-1}\right)}$$

where n is the number of data values (sample size) and $\bar{x}$ is the mean.

1G The five-number summary and the boxplot

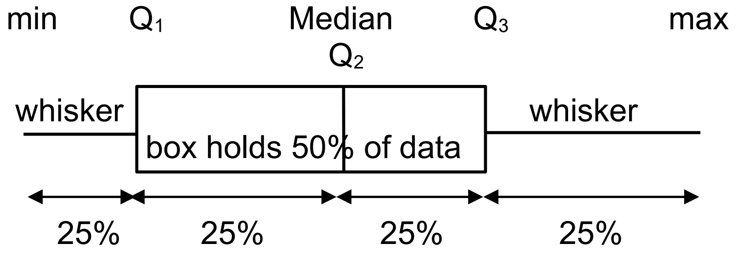
Five-number summary

of data arranged in ascending order.

- Minimum value (min)
- Lower quartile (Q_1) : the median of the bottom half. 25% of the data is below this number.
- Median (Q_2) : the median value. 50% of the data is above this number, 50% below
- Upper quartile (Q_3) : the median of the top half. 75% of the data is below this number.
- Maximum value (max)

Boxplot

is the five-number summary in picture form.



Outliers

Outlier : any data point

smaller than $Q_1 - 1.5 \times IQR$ (lower fence)

or

bigger than $Q_3 + 1.5 \times IQR$ (upper fence)

Shown with a dot or a cross on a boxplot.

Worked example : boxplot with outlier

33 students spent these hours on a school project.

2 3 4 9 9 13 19 24 27 35 36 37 40 48 56 59 71 76 86 90

92 97 102 102 108 111 146 147 147 166 181 226 264

Odd number of data points. Median value at point $\frac{33+1}{2}$ which is 71.

Divide data into two groups. 71 omitted from both groups.

[2 3 4 9 9 13 19 24 27 35 36 37 40 48 56 59]

16 = even number of data points. Q_1 is average of values at $\frac{16}{2}$ and $\frac{16}{2}+1$

$Q_1 = \frac{24+27}{2} = 25.5$

[76 86 90 92 97 102 102 108 111 146 147 147 166 181 226 264]

Q_3 is average of values at $\frac{16}{2}$ and $\frac{16}{2}+1$

$Q_3 = \frac{108+111}{2} = 109.5$

IQR = $Q_3 - Q_1$ = 109.5 – 25.5 = 84

Lower fence = Q_1 – 1.5 × IQR

= 25.5 – 1.5 × 84

= – 100.5

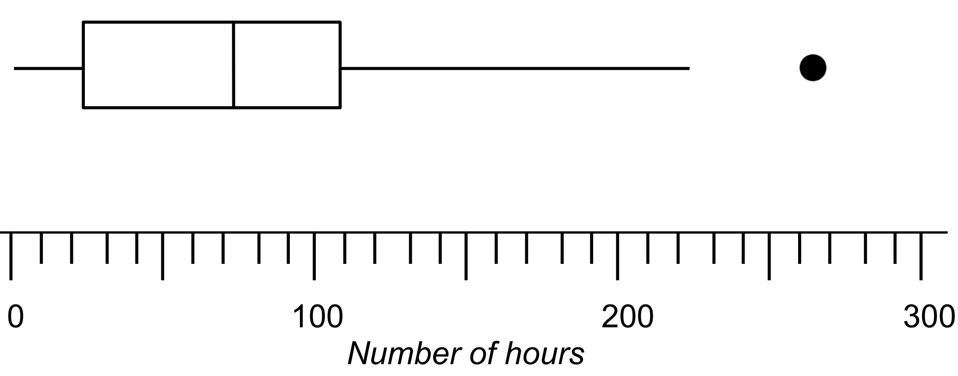
Upper fence = Q_3 + 1.5 × IQR

= 109.5 + 1.5 × 84

= 235.5

264 is above the upper fence, so it is an outlier and will be drawn with a dot.

The whisker will extend to 226, which is the largest value that is not an outlier.



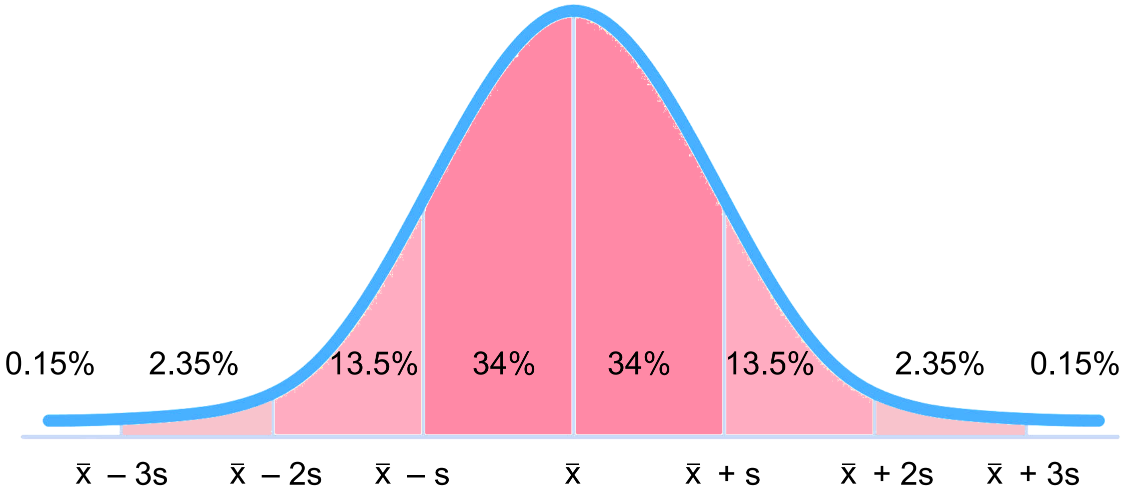
Report Boxplot

The distribution is / is not symmetric and without outliers / but with outliers. The distribution is centred at BLAH, the median value. The spread of the distribution, as measured by the IQR, is BLAH and, as measured by the range, BLAH. There are BLAH outliers: BLAH, BLAH, BLAH and BLAH.

1H The normal distribution and the 68–95–99.7% rule

For any data distribution which is approximately symmetric and bell shaped, approximately:

68% lie within $\bar{x} \pm s$ 95% lie within $\bar{x} \pm 2s$ 99.7% lie within $\bar{x} \pm 3s$



Example : find % from interval

The distribution of delivery times for pizzas made by House of Pizza is approximately normal, with a mean of 25 minutes and a standard deviation of 5 minutes.

a) What percentage of pizzas have delivery times of between 15 and 35 minutes?

b) What percentage of pizzas have delivery times of greater than 30 minutes?

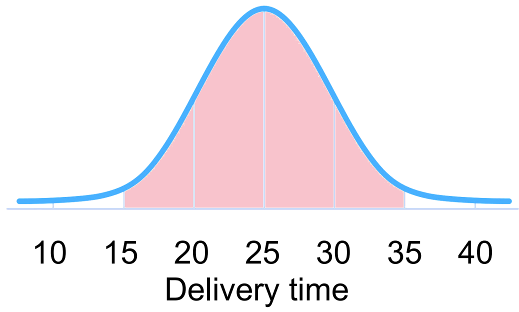
c) In 1 month, House of Pizza delivers 2000 pizzas. Approximately how many of these pizzas are delivered in less than 10 minutes?

a)

Mean is the number in the middle.

Add the standard deviation, that's 5, at each mark to the right, and subtract the standard deviation at each mark to the left.

Shade the region under the normal curve representing delivery times of between 15 and 35 minutes.



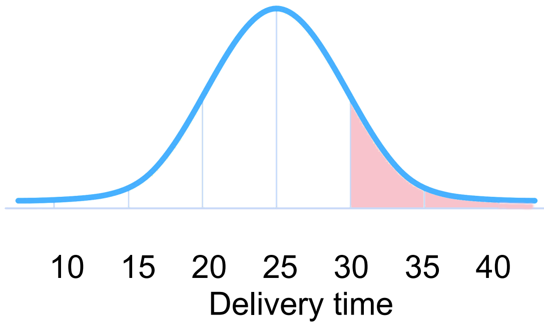
Refering to the labelled bell curve above, the shaded areas represent

95% of pizzas have a delivery time of between 15 and 35 minutes.

13.5% + 34% + 34% + 13.5% = 95%

b)

Shade the region under the normal curve representing delivery times of greater than 30 minutes.



13.% + 2.35% + 0.15% = 16%

c)

2000 pizzas delivered.

0.15% delivered in less than 10 minutes.

0.15% × 2000 = 3

Example : find interval from %

The distribution of the diameter of bolts is approximately symmetric and bell shaped, with a mean of 5 mm and standard deviation of 0.01mm.

If approximately 68% of the bolts measure between a and b, what are possible values for a and b?

The interval which contains 68% of the bolts is 1SD either side of the mean.

a = 5 - 0.01 = 4.99 mm   b = 5 + 0.01 = 5.01 mm

Standard score, z-score

$$\text{standard score} = \frac{\text{actual score} - \text{mean}}{\text{standard deviation}}$$

$$z = \frac{x - \bar{x}}{s}$$

Standard score is a measure of how average you are. A score of 0 means you are exactly the same as expected, average. A score of less than 0, a negative number, means you are somewhat lacking, below average. A score of more than 0, a positive number, means you are a standout performer, above average.

Example : raw score to z-score

IQ is normally distributed, with a mean, $\bar{x}$, of 100 and a standard deviation, s, of 15.

What is the z-score for a student who gets 112 in an IQ test? What does it mean?

$$z = \frac{x - \bar{x}}{s} = \frac{115 - 100}{15} = 0.8$$

This student is less than one standard deviation away from average. They are nothing remarkable.

Example : z-score to raw score

A student achieves a z-score of 2.2. What did they score in the IQ test? Which percentage do they fit into?

$$x = z \times s + \bar{x}$$

$$\text{score} = 2.2 \times 15 + 100$$

Student's test score = 133.

z-score of 2.2 means they are between 2 and 3 standard deviations above the average. They are among the top 2.5% of the population.

3 INVESTIGATING LINEAR ASSOCIATIONS

3A Least squares regression line applied to numerical data

is given by, $y = a + bx$, where:

the slope (b) is given by: $b = \frac{rs_y}{s_x}$

and the intercept (a) is then given by : $a = \bar{y} - b\bar{x}$

- r is the correlation coefficient
- s_x and s_y are the standard deviations of x and y
- $\bar{x}$ and $\bar{y}$ are the mean values of x and y.

The least squares line

minimises the sum of the squares of the residuals.

Assumptions for fitting a least squares line to data are the same as for using the correlation coefficient, r. These are that:

- the data is numerical
- the association is linear
- there are no clear outliers

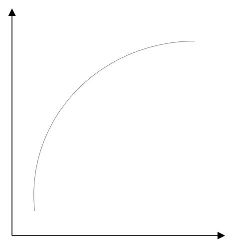
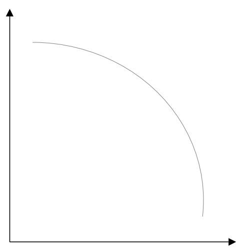
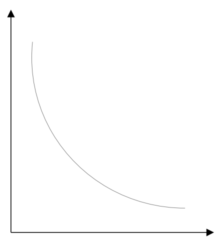
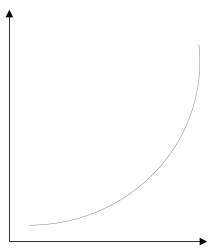
4 : DATA TRANSFORMATION

Turn a bendy curve straight,

and then linear regression an equation for the straight line.
As if by magic, that equation works for the bendy curve we started with.

This stuff is done on the calculator.

6 Transformations, 4 graph shapes

Possible transformations	Graph shape		Possible transformations
y^2 $\log(x)$ $\frac{1}{x}$			y^2 x^2
$\log(y)$ $\frac{1}{y}$ $\log(x)$ $\frac{1}{x}$			$\log(y)$ $\frac{1}{y}$ x^2

The best transformation

is the one which results in the best linear model.

For each transformation, consider :

■ The residual plot, in order to evaluate the linearity of the transformed association.

■ The value of the coefficient of determination, r^2.

5B Smoothing with moving means

Smoothing can sometimes remove some of the fluctuations in time series data so that we can see the underlying trend.

Three-moving mean

replace each data value with the mean of that value and the one on each side.

smoothed $y_2 = \dfrac{y_1+y_2+y_3}{3}$

The first and last points in the time series do not have values on each side, so they are omitted.

Centring

involves taking a two-moving mean of the already smoothed values so that they line up with the original data values. Smoothing with centring is only required when smoothing using an even number of data values.

Example : the table below gives the temperature (◦C) recorded at a weather station at 9.00 a.m. eachday for a week.

Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Temperature	18.1	24.8	26.4	13.9	12.7	14.2	24.9

a) Calculate the four-moving mean smoothed temperature with centring for Thursday.

1 For four-mean smoothing with centring, write down the five data values centred on Thursday

24.8   26.4   13.9   12.7   14.2

2 Calculate the mean of the first four values (mean 1)

$$\text{mean 1} = \frac{24.8+26.4+13.9+12.7}{4} = 19.45$$

and the mean of the last four values (mean 2).

$$\text{mean 2} = \frac{26.4+13.9+12.7+14.2}{4} = 16.8$$

The centred mean is then the average of mean 1 and mean 2

$$\text{centred mean} = \frac{\text{mean}\,1+\text{mean}\,2}{2}$$

$$= \frac{19.45+16.8}{2} = 18.125$$

4 Write down your answer

The four-mean smoothed temperature centred on Thursday is 18.1 °C (to 1 d.p.).

8 REDUCING BALANCE LOANS, ANNUITIES AND INVESTMENTS

8A Compound interest investments with additions to the principal

V_0 = the Principal, $R = 1 + \dfrac{r}{100 \times PpY}$ D = additional payment

V_n = the value of the investment after n payments

$V_{n+1} = RV_n + D$

Finance solver

■ PV : Negative: you make an investment by giving the bank some money.

■ Pmt : Negative: you make regular payments to the bank.

■ FV : Positive: after the payment is made and the investment matures, the bank will give you the money.

10E Matrix multiplication

number of **columns in 1st** matrix must be **same as** number of **rows in 2nd** matrix.	$(m \times n)\ (n \times p) = (m \times p)$ must be the same
product matrix has number of **rows** from **1st** matrix number of **columns** from **2nd** matrix.	$(m \times n)\quad (n \times p) \quad = \quad (m \times p)$ this many columns this many rows

Multiply each row in A by each column in B. (Run along the row, dive into the column.)

Multiply row m by column p and the result goes in a_{mp} of the product matrix.

Multiply row 1 by column 1 and the result goes in a_{11} of the product matrix.

Multiply row 2 by column 1 the result goes in a_{21} of the product matrix.

Let $A = \begin{bmatrix} 1 & -2 \\ 2 & 0 \end{bmatrix}$ $B = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$ Then $AB = \begin{bmatrix} 1\times4+(-2)\times1 \\ (-2)\times4+0\times1 \end{bmatrix} = \begin{bmatrix} 2 \\ -8 \end{bmatrix}$

Let $A = \begin{bmatrix} 2 & 3 \\ 5 & 4 \\ 1 & 6 \end{bmatrix}$ $B = \begin{bmatrix} 7 & 10 \\ 9 & 8 \end{bmatrix}$ Then $AB = \begin{bmatrix} 2\times7+3\times9 & 2\times10+3\times8 \\ 5\times7+4\times9 & 5\times10+4\times8 \\ 1\times7+6\times9 & 1\times10+6\times8 \end{bmatrix} = \begin{bmatrix} 41 & 44 \\ 71 & 82 \\ 61 & 58 \end{bmatrix}$

Matrix powers

We define the various powers of matrices as:
A^2 as A × A,
A^3 as A × A × A,
A^4 as A × A × A × A and so on.

Only square matrices can be raised to a power.

11 TRANSITION MATRICES AND LESLIE MATRICES

11A Setting up a transition matrix

Transition means changing from one state to another.

Transition matrix, T, represents the amount of change occuring at each step.

Used with a state matrix, S_n, which lists the numbers in each state at step n.

Essential knowledge.

A transition matrix is always a square matrix.

Each column total of the proportions (or percentages) must equal 1 (100%).

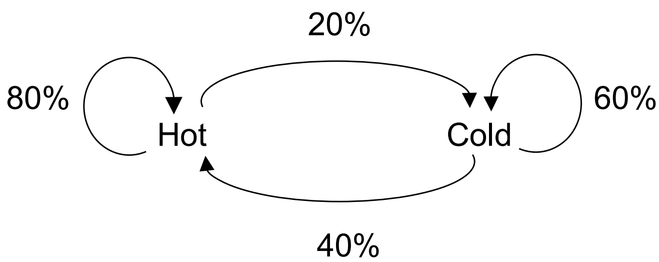
Example : a cafe notices that

■ 80% of customers who buy hot food (H) for lunch will buy hot food again the next day.

■ 60% of customers who buy cold food (C) for lunch will buy cold food again the next day.

20% of customers who buy hot food (H) for lunch will buy cold food the next day.

40% of customers who buy cold food (C) for lunch will buy hot food the next day.



Circles go on the leading diagonal.

Hot to Cold is at a_{21}

Write percentages as decimals.

Day 1

$$\text{Day 2}\quad \begin{array}{c} \begin{matrix} H & C \end{matrix} \\ \begin{bmatrix} 0.8 & 0.4 \\ 0.2 & 0.6 \end{bmatrix} \end{array} \begin{matrix} H \\ C \end{matrix}$$

13C Walks, trails, paths, circuits and cycles

Walk

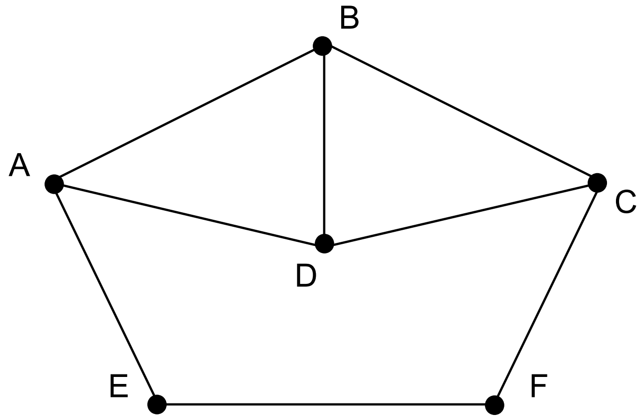
is a sequence of edges linking successive vertices, that connects two different vertices in a graph.

An example of a walk is

A – D – C – B – D – A – E

Note1: A – D is walked along in both directions.

Note2: A walk does not require all of its edges or vertices to be different.



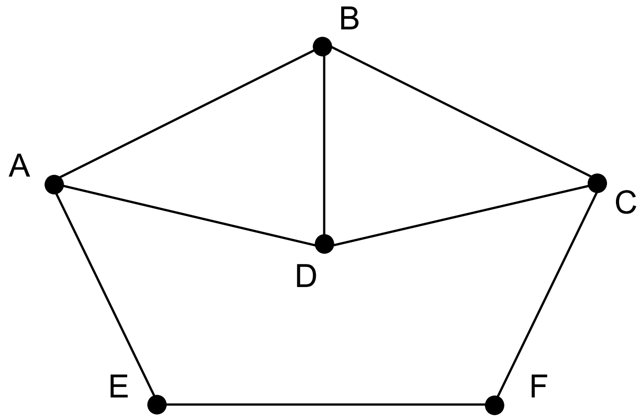
Trail

is a walk with no repeated edges.

An example of a trail is

A – D – C – B – A – E – F – C

There are no repeated edges. But there are two repeated vertices, A and C. This is permitted on a trail.



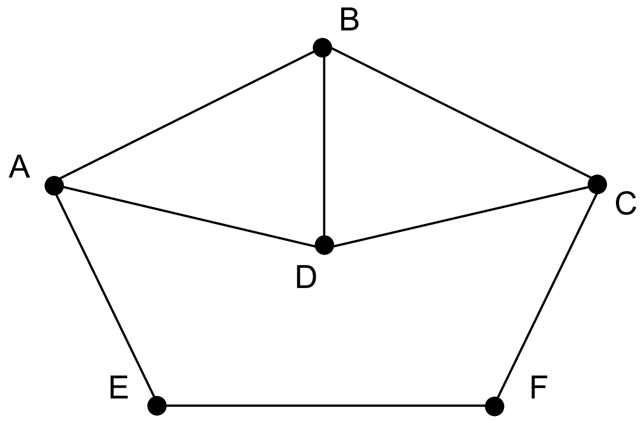
Path

is a walk with no repeated edges and no repeated vertices.

An example of a path is

A – D – B – C – F – E

There are no repeated edges and no repeated vertices.



14B Matching and allocation problems

The Hungarian algorithm

allocates people to roles

This is the cost matrix for 4 machines, A – D, that can be operated by 4 people, 1 – 4.

Person 2 takes 30 minutes to complete the task on machine B.

The Hungarian algorithm allocates people to machines to minimise the cost.

	Role	A	B	C	D
Person	1	30	40	50	60
	2	70	30	40	70
	3	60	50	60	30
	4	20	80	50	70

Step 1:

Subtract the lowest value in each row, from every value in that row.

	Role	A	B	C	D
Person	1	0	10	20	30
	2	40	0	10	40
	3	30	20	30	0
	4	0	60	30	50

Step 2:

If the minimum number of lines required to cover all the zeros in the table is equal to the number of allocations to be made, jump to step 6.

Otherwise,continue to step 3.

	Role	A	B	C	D
Person	1	0	10	20	30
	2	40	0	10	40
	3	30	20	30	0
	4	0	60	30	50

Step 3:

If a column does not contain a zero, subtract the lowest value in that column from every value in that column.

	Role	A	B	C	D
Person	1	0	10	10	30
	2	40	0	0	40
	3	30	20	20	0
	4	0	60	20	50

Step 4:

If the minimum number of lines required to cover all the zeros in the table is equal to the number of allocations to be made, jump to step 6.

Otherwise, continue to step 5a.

	Role	A	B	C	D
Person	1	0	10	10	30
	2	40	0	0	40
	3	30	20	20	0
	4	0	60	20	50

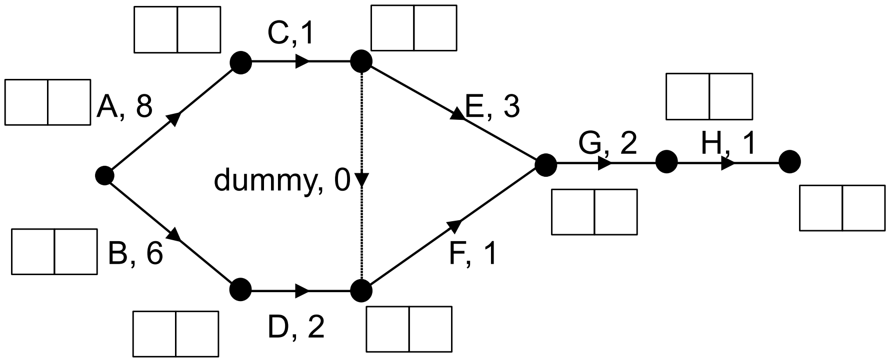
Earliest starting times (EST)

are found by forward scanning.

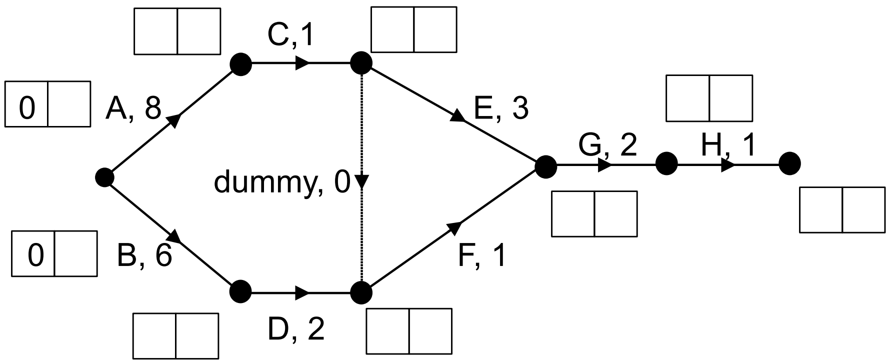
Forward scanning

1 Draw a double box next to each vertex.

If more than one activity begins at a vertex, draw a box for each of these activities.



2 Activities that begin at the start of the project have an EST of zero (0).



3 Calculate the EST of each activity of the project by adding the EST of the immediate predecessor to the duration of the immediate predecessor.