Displaying Data

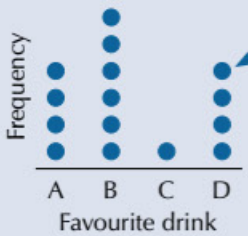
Data can be shown on lots of different charts and graphs. The ones you need to know for this course are shown on the next few pages. You've probably seen a fair few of them before, so they should be bread and butter by now.

Data can be represented Graphically

Bar charts, **dot plots** and **vertical line charts** are all simple ways to show how data is distributed.

15 people were given a blind taste test of four drinks, labelled A, B, C and D, then asked which was their favourite. The number of people who voted for each drink is shown on the diagrams below.

Dot plot:



Each dot represents 1 vote, so 4 people voted for D.

For a large amount of data, drawing dots would be inconvenient, so a bar chart might be more appropriate.

Bar chart:

Vertical line charts are like bar charts, but using lines instead of bars (as you might expect).

Stem and Leaf Diagrams show all the data

Stem and leaf diagrams are another way to represent data — the diagrams show the data values themselves. Each data value is split into a '**stem**' and a '**leaf**'. A complete stem and leaf diagram looks something like this:

This stem and leaf diagram shows the ages of readers of the Daily Pry newspaper.

These are the **stems**. These are the **leaves**.

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

This is the **key**.

Key: 2 | 3 means 23

A stem and leaf diagram always needs a **key** to tell you how to read it. So, in the stem and leaf diagram above, the **first row** represents the values **22**, **23**, **24**, **27**, **27**, and **28**, while the **second row** represents the values **30**, **30**, **32**, **35** and **36**. You can read the other two rows in a similar way.

You can show Two Data Sets on a Back-To-Back stem and leaf diagram

Two stem and leaf diagrams can be drawn either side of the same stem — i.e. **back-to-back**.
The data on the left hand side of the stem is read '**backwards**' — because the stems are on the right of the leaves.

Example: a) Draw a back-to-back stem and leaf diagram to represent the following data:
Boys' test marks: 50, 20, 18, 38, 34, 19, 8, 44, 15, 32, 9, 19, 41, 26, 22
Girls' test marks: 36, 24, 42, 46, 35, 12, 38, 45, 31, 38, 21, 43, 37, 27, 29, 46

1) Use the tens digits as the stems.

2) Add the units digits one at a time as the leaves.

3) Put the leaves in order.

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

Don't forget about the key.

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

4) Add a key.

Key 0|2|1 means 20 for boys and 21 for girls

b) Compare the shapes of the two distributions.

The boys' marks show **more variation** (larger range) and are **more evenly distributed**.
The girls' marks are generally **higher** than the boys' marks.

It might help to picture the stem and leaf diagram as back-to-back bar charts.

Displaying Data

From a stem and leaf diagram you could be asked to find measures of central tendency (e.g. mean, median, mode).

Example: For the test marks data on the previous page, find:
a) the mode of the boys' marks,
b) the median of the boys' marks and the median of the girls' marks. Compare your answers.

a) All the data values for the boys' data appear once except 19, which appears twice — so **mode = 19**.

b) boys' data: there are 15 values, so the median is the 8th value — so **median = 22**.

girls' data: there are 16 values, so the median is halfway between the 8th and 9th values
— so **median** = (36 + 37) ÷ 2 = **36.5**

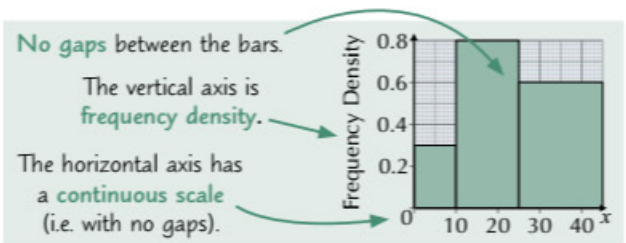
The median of the girls' data is **higher** than the median of the boys' data, which suggests that girls scored higher in the test than boys.

Histograms show Frequency Density

Histograms are glorified bar charts. The main difference is that you plot the **frequency density** rather than the frequency. Frequency density = **frequency** ÷ **class width**.

The **area** of a bar (**not** its height) represents the **frequency**.

To get histograms right, you have to use the right **upper and lower boundaries** to find each class width (see p.146).



Practice Questions

Q1 a) Draw a vertical line chart for the data below:

Rating (1-5)	1	2	3	4	5
Frequency	2	4	9	5	1

b) Find the mode of the data.

c) Comment on the distribution of the data.


Vinnie was a talented stem and leaf drawer.

Q2 Draw a stem and leaf diagram to represent the data below:
Class attendance (%): 89, 92, 90, 95, 100, 85, 77, 87, 95, 98

Q3 Find the frequency density in the following cases:
a) frequency = 25, class width = 10, b) frequency = 33, class width = 15.

Exam Questions

Q1 The number of runs scored by two cricketers, in 10 matches, are shown below.
Cricketer A: 50, 32, 17, 45, 0, 26, 3, 50, 15, 12
Cricketer B: 27, 22, 33, 34, 38, 44, 41, 17, 20, 31

a) Draw a back-to-back stem and leaf diagram to represent the data. [2 marks]
b) Find the median number of runs scored by both cricketers. [2 marks]

Q2 The stem and leaf diagram shows the ages at which 30 men and 16 women became grandparents.

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

Key: 5 | 6 | 2 means a man who became a grandfather at 65 and a woman who became a grandmother at 62.

a) Find the median age for the men. [1 mark]
b) Compare the distribution of the two data sets. [2 marks]

Time to make like a stem and leaf diagram and leave...

...but read these last few notices first. You can tell a lot about central tendency and variation from the shape of a distribution. And don't worry if you're a bit unsure on histograms, they're covered in more detail on the next page.

Cumulative Frequency Graphs and Boxplots

Cumulative frequency means 'running total'. Cumulative frequency graphs make medians and quartiles easy to find.

Use **Cumulative Frequency Graphs** to estimate the **Median** and **Quartiles**

Example: The ages of 200 students are shown in the table.

Age in completed years	11-12	13-14	15-16	17-18
Number of students	50	65	58	27

Draw a cumulative frequency graph and use it to estimate the median age, the interquartile range of ages, and how many students have already had their 18th birthday.

1) First draw a table showing the **upper class boundaries** and the **cumulative frequency** (CF):

Age in completed years	Upper class boundary (ucb)	Number of students, f	Cumulative frequency (CF)
Under 11	11	0	0
11-12	13	50	50
13-14	15	65	115
15-16	17	58	173
17-18	19	27	200

The **first** reading in a **cumulative frequency** table **must** be **zero** — so add this **extra row** to show the number of students with age **less than 11** is 0.

The CF is the number of students with an age **up to** the ucb — it's basically a **running total**.

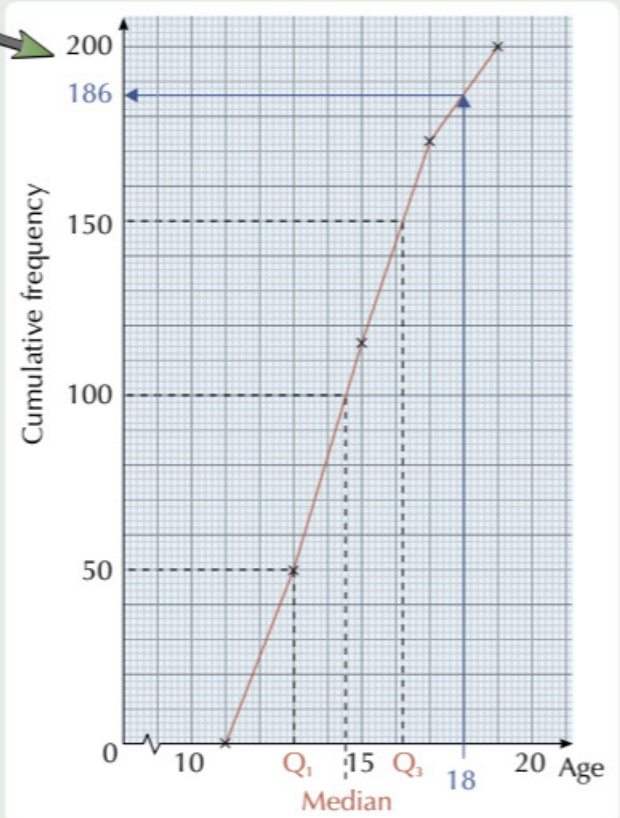
The **last** number in the CF column should always be the **total number** of readings.

People say they're '18' right up until their 19th birthday — so the **ucb** of class 17-18 is **19**.

Always plot the upper class boundary of each class.

Next draw the **axes** — cumulative frequency **always** goes on the **vertical axis**. Here, age goes on the other axis.

Then plot the **upper class boundaries** against the **cumulative frequencies**, and join the points.



2) To estimate the **median** from a cumulative frequency graph, go to the **median position** on the vertical scale and read off the value from the horizontal axis.

median position $= \frac{1}{2} \times 200 = 100$,

so median ≈ **14.5 years**

You can only **estimate** the median, since your data values are in **groups**. This is similar to linear interpolation — see page 147.

Then you can estimate the **quartiles** in the same way. Find their positions first:

Q_1 position $= \frac{1}{4} \times 200 = 50$,
so lower quartile, Q_1 ≈ **13 years**

Q_3 position $= \frac{3}{4} \times 200 = 150$,
so upper quartile, Q_3 ≈ **16.2 years**

IQR $= Q_3 - Q_1 = 16.2 - 13 =$ **3.2 years**

Because the question says estimate, a **range** of answers would be **correct** for the median and IQR — e.g. anything between 14.25 and 14.75 for the median and anything between 3 and 3.5 for the IQR.

3) To estimate how many students have **not** yet had their 18th birthday, go up from 18 on the **horizontal axis**, and read off the number of students '**younger**' than 18 (= 186).

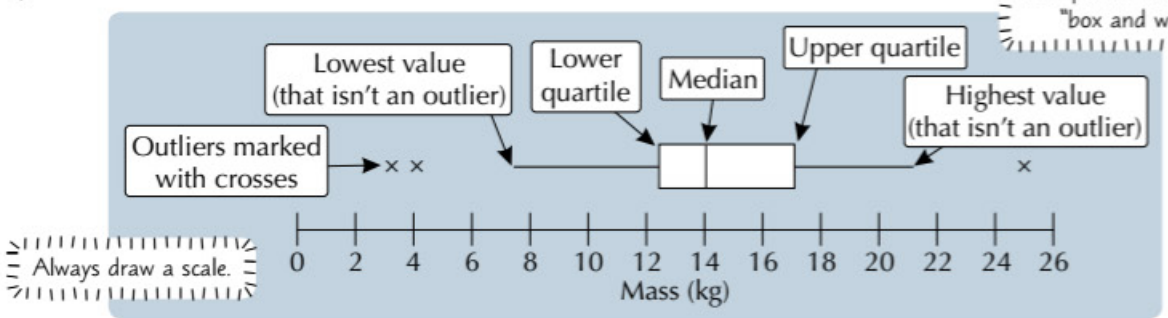
Then the number of students who are 'older' than 18 is approximately 200 – 186 = **14**.

Cumulative Frequency Graphs and Boxplots

Box Plots are a Visual Summary of a Distribution

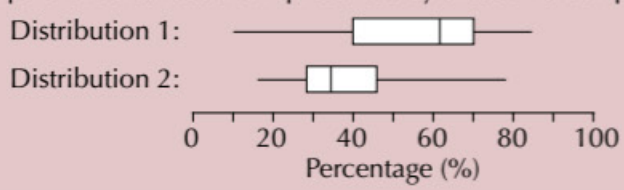
Box plots show the median and quartiles in an easy-to-look-at kind of way.
They look like this:

Box plots are sometimes called "box and whisker plots".



Use Box Plots to Compare two Distributions

Example: Compare the distributions represented by these two box plots:



Location: The **median** is higher for Distribution 1, showing that the data values are **generally higher** than for Distribution 2.

Variation: The **interquartile range** (IQR) and the **range** for Distribution 1 are higher, showing that the values are **more varied** for Distribution 1 than for Distribution 2.

Tian Tian was plotting something fiendish. It was a box plot...

Practice Questions

Q1 Draw a cumulative frequency diagram of the data given in this table.

Length of tadpole (cm)	0-2	2-4	4-6	6-8
Number of tadpoles	22	12	4	2

Use your diagram to estimate the median and interquartile range.

Q2 Draw a box and whisker diagram for the data below, using the fences 1.5 × IQR above Q_3 or below Q_1 to identify any outliers. Amount of pocket money (in £) received per week by twenty 15-year-olds:
10, 5, 20, 50, 5, 1, 6, 5, 15, 20, 5, 7, 5, 10, 12, 4, 8, 6, 7, 30.

Exam Question

Q1 Two workers iron clothes. Each irons 10 items, and records the time it takes for each, to the nearest minute:

Worker A: 3, 5, 2, 7, 10, 4, 5, 5, 4, 12
Worker B: 3, 4, 8, 6, 7, 8, 9, 10, 11, 9

a) Find the median, the lower quartile and the upper quartile for worker A's times. [2 marks]

b) On graph paper, draw two box plots to show this data, one for each worker.
Use the same scale for both plots and assume there are no outliers. [4 marks]

c) Worker A claims he deserves a pay rise because he works faster than Worker B.
State, giving a reason, whether the data given above supports Worker A's claim. [2 marks]

"It's a cumulative frequency table," she said. It was all starting to add up...

'Cumulative frequency' sounds a bit scarier than 'running total' — but just remember, they're the same thing. And remember to plot the points at the upper class boundary — this makes sense if you remember that a cumulative frequency graph shows how many data values are less than the figure on the x-axis. The rest is more or less easyish.

Interquartile Range and Outliers

Outliers fall Outside Fences

An **outlier** is a **freak** piece of data that lies a long way from the rest of the readings. There are various ways to decide if a reading is an outlier — the method to use will depend on what information you're given in the question.

Example: A data value is said to be an outlier if it is more than 1.5 times the IQR above the upper quartile or more than 1.5 times the IQR below the lower quartile. The lower and upper quartiles of a data set are 70 and 100. Decide whether the data values 30 and 210 are outliers.

This is one of two methods for finding outliers that you should be familiar with. The other one is 'more than 2 standard deviations away from the mean'.

First you need the IQR: $Q_3 - Q_1 = 100 - 70 = 30$

Then it's a piece of cake to find where your **fences** are. 25 and 145 are called **fences**. Any reading lying **outside** the fences is considered an **outlier**.

Lower fence: $Q_1 - (1.5 \times IQR) = 70 - (1.5 \times 30) = 25$ Upper fence: $Q_3 + (1.5 \times IQR) = 100 + (1.5 \times 30) = 145$

30 is **inside** the lower fence, so it is **not** an outlier. 210 is **outside** the upper fence, so it **is** an outlier.

Outliers Affect what Measure of Variation is Best to Use

- Outliers affect whether the **variance** and **standard deviation** are good measures of **variation**.
- Outliers can make the variance (and standard deviation) **much** larger than it would be otherwise — which means these **freak** pieces of data are having more influence than they deserve.
- If a data set contains outliers, then a better measure of variation is the **interquartile range**.

Use Central Tendency and Variation to Compare Distributions

Example: The table below summarises the marks obtained in Maths 'calculator' and 'non-calculator' papers. Comment on the location and variation of the distributions.

	Lower quartile, Q_1	Median, Q_2	Upper quartile, Q_3	Mean	Standard deviation
Calculator Paper	40	58	70	55	21.2
Non-calculator Paper	35	42	56	46.1	17.8

Location: The **mean**, the **median** and the **quartiles** are all higher for the calculator paper. This means that scores were **generally higher** on the calculator paper.

Variation: The **interquartile range** (IQR) for the calculator paper is $Q_3 - Q_1 = 70 - 40 = 30$.
The **interquartile range** (IQR) for the non-calculator paper is $Q_3 - Q_1 = 56 - 35 = 21$.
So the **IQR** and the **standard deviation** are both **higher** for the calculator paper.
So the scores on the calculator paper are **more spread out** than for the non-calculator paper.

Practice Question

Q1 A data value is considered to be an outlier if it's more than 2 times the standard deviation above or below the mean. If the mean and standard deviation of a data set are 72 and 6.7 respectively, decide which of the following data values are outliers: a) 85 b) 95 c) 0

Exam Question

Q1 The table shows the number of hits received by people at a paintball party.

No. of hits	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Frequency	2	4	6	7	6	4	4	2	1	1	0	0	0	1

a) Find the median and mode number of hits. [3 marks]

b) An outlier is a data value which is more than $1.5 \times (Q_3 - Q_1)$ above Q_3 or below Q_1.
Is 25 an outlier? Show your working. [2 marks]

c) Explain why the median might be considered a more reliable measure of central tendency than the mean for a data set that is thought to contain an outlier. [1 mark]

I like my data how I like next door's dog — on the right side of the fence...

Measures of location and variation are supposed to capture the essential characteristics of a data set in just one or two numbers. Don't choose an average that's heavily affected by freaky, far-flung outliers — it won't be much good.