

KEY WORDS:

- | | |
|-----------------|------------|
| bias | sample |
| continuous data | stratified |
| discrete data | systematic |
| experiment | survey |
| primary | two-way |
| questionnaire | table |
| random | |

- **Primary** data is data that you collect yourself. You can collect data from a **survey** or an **experiment**.
- **Secondary** data is data taken from an existing source such as newspapers or the internet.
- Data can be **discrete** or **continuous**.
- A **two-way table** shows two sets of data about the same group of people or things, for example, hair colour and eye colour.

You can count discrete data, for example the number of eggs in a nest. Continuous data is data you measure, for example weight, height.

	Hair colour
Eye colour	

You need to choose a **sample** that will not be biased, for example **bias** can happen when you only ask your friends or only ask people in a certain age group. In general the larger the sample size the more reliable the data analysis.

To choose a **random sample**, assign a number to each possible member of the group and use a random number generator to pick numbers.

To find a 20% **systematic sample** start from a random value, say the 3rd value and then choose every 5th ($20\% = \frac{1}{5}$) value, that is 3rd, 8th, 13th, 18th, 23rd ...

In all types of sampling first assign a number to each possible data value.

In a **stratified sample** divide the population into groups and choose a random sample from each group in the ratio of their sizes.

- **Questionnaires** should be relevant and useful to your survey.

Use clear language. Cover all options and leave no gaps between the answer boxes.

EXAMPLE

This question appeared in a survey about time spent using a mobile phone.
 How much time do you spend on your mobile phone?
 Less than 5 minutes Up to 10 minutes Over an hour
 Criticise the question, and write a better question.

The question does not have a time frame.
 It does not specify just making phone calls, or also listening to music and playing games and so on.
 The answer choices have gaps and overlaps.
 How long, on average, do you spend making calls on your mobile each day?
 Less than 10 minutes 10–30 minutes Over 30 minutes

A machine produces buttons.
Three workers use the machine and produce the following numbers of buttons.

Abi – 5890 Ben – 4140 Carly – 6770

Their supervisor wants to take a stratified sample of 400 of these buttons.

Work out how many she should sample from Abi, Ben and Carly.

Work out the total. Find the fraction of 400 for each person. Check that the total of your three fractions is 400.

Total produced $5890 + 4140 + 6770 = 16\,800$

Abi $400 \times \frac{5890}{16800} = 140.2$ so 140 buttons

Ben $400 \times \frac{4140}{16800} = 98.6$ so 99 buttons

Carly $400 \times \frac{6770}{16800} = 161.2$ so 161 buttons

Check $140 + 99 + 161 = 400$

Exercise S1

- 1 Design an observation sheet to collect data on colour and type of vehicle in a road traffic survey.
- 2
 - a Design an observation sheet to collect data on women's shoe sizes and glove sizes. You may assume shoe sizes range from 3 to 8 and glove sizes range from extra small to large.
 - b A survey gave the following information.
 - 7 women had shoe size 5 and glove size medium
 - 9 women had shoe size 4 and glove size small
 Add this information to your observation sheet
- 3 Answers to a crossword puzzle are completed either across or down. In a crossword puzzle book all the crossword puzzles have across answers of 4 to 9 letters long and down answers of 5 to 10 letters long.
 - i Design an observation sheet to capture data about the lengths of answers in the book.
 - ii There were 12 crosswords with across answers 7 letters long and down answers 6 letters long. Write this value in the correct position on your observation sheet.
 - b Debbie wants to do a survey of the crosswords in the book. She gives each of twelve work colleagues a copy of the crossword book and a questionnaire to complete. Give two reasons why Debbie's survey may be biased.

- An **average** is a representative value of a set of data.
- The **mode** is the data value that occurs most often.
- In grouped data the **modal class** is the class with the highest frequency.
- The **median** is the middle number when data are arranged in order.
- In grouped data you can find the class in which the median lies but not an actual median value.
- The **mean** is calculated by adding all the data values then dividing by the number of pieces of data.
- In grouped data you calculate an **estimated mean** using the midpoint and the frequency of each class.

KEY WORDS:

average	median
estimated mean	modal class
mean	mode

If there are two middle numbers the median is the middle value of these two, for example, in the data set 2 2 4 5 7 8 10 11 the middle values are 5 and 7 so the median = 6.

EXAMPLE

On boxes of drawing pins the average contents are labelled as 50. The table shows the numbers of drawing pins in 35 boxes.

Drawing pins	48	49	50	51	52	53	54
Frequency	3	6	21	1	2	1	1

Joni says that the average is 50, whatever measure of average is used. Explain why Joni is correct.

The mode is 50, as 50 has the highest frequency (21).
 For 35 boxes, the middle value is $\frac{1}{2}(35 + 1) = 18$ th value,
 $3 + 6 = 9, 9 + 21 = 30$
 So the 18th value, the median = 50
 Total number of pins = $48 \times 3 + 49 \times 6 + 50 \times 21 + 51 + 52 \times 2 + 53 + 54$
 $= 1750$
 mean = $1750 \div 35 = 50$
 All three types of average give an answer of 50, Joni is correct

To find the median in a small data set find the $\frac{1}{2}(n + 1)$ th value.
 Imagine listing the numbers of pins in order - 50 would be the 18th number in the list.

The table summarises the number of miles Reuben cycled each day for the first 29 days in April.

Miles cycled (m)	$0 \leq m < 20$	$20 \leq m < 40$	$40 \leq m < 60$	$60 \leq m < 80$
Frequency	5	9	11	4

- Work out an estimate of the mean number of miles Reuben cycled.
- How many miles could Reuben cycle on April 30th so that the class interval in which the median lies does not change? Explain your answer.

- a** Total number of miles = $10 \times 5 + 30 \times 9 + 50 \times 11 + 70 \times 4 = 1150$
 $1150 \div 29 = 39.655\dots$ so the mean = 39.7 miles

- b** $\frac{1}{2}(29 + 1) = 15$ th value $5 + 9 = 14$, so 15th value, the median, is in the class $40 \leq m < 60$
 With 30 values median will be $\frac{1}{2}(30 + 1) = 15.5$ th value,
 so you average the 15th and 16th values. Both of these must be in the class $40 \leq m < 60$ for the median to stay where it is.
 So on April 30th Reuben must cycle a distance of 40 miles or more.

Estimate does not mean guess. Use the frequency and the midpoint of each class. Remember to divide by the total frequency not the number of classes.

Exercise S2

- 1 Henry drew this table to show the number of tracks on each of the CDs he owned.

Number of tracks	7	8	9	10	11
Number of CDs	4	6	12	5	2

- Write down
 - the median
 - the mode number of tracks.
 - Work out the mean number of tracks.
- 2 The numbers and prices of theatre tickets available for a performance are shown in the table.

Ticket price, £	£50	£40	£25	£15
Number available	140	50	200	10

- Calculate the mean price of a ticket.
 - For a special performance one evening all ticket prices were reduced by £5. What was the mean price of a ticket for that performance?
- 3 The numbers of hats of different hat sizes sold by a department store are given in the table.

Hat size	$6\frac{3}{8}$	$6\frac{1}{2}$	$6\frac{5}{8}$	$6\frac{3}{4}$	$6\frac{7}{8}$	7	$7\frac{1}{8}$	$7\frac{1}{4}$	$7\frac{3}{8}$	$7\frac{1}{2}$	$7\frac{5}{8}$
Number sold	1	0	2	3	8	15	12	3	0	0	0

When the manager of the store orders new hats to sell which type of average should he use? Give reasons for your answer and show your calculations.

Overmatter exercise S2

- 4 The table shows the times a sample of 84 students spent doing homework one evening.

Time $m(\text{min})$	$0 < m \leq 20$	$20 < m \leq 40$	$40 < m \leq 60$	$60 < m \leq 80$	$80 < m \leq 100$	$100 < m \leq 120$
Frequency	1	21	11	22	8	21

- a Work out an estimate of the mean time.
b Which class contains the median?
c Explain why the modal class may not be a good measure of average to use with these data.
- 5 Ahmed recorded the number of hours of sunshine in May and June.

May

Number of hours sunshine h	Number of days
$0 \leq h < 2$	5
$2 \leq h < 4$	14
$4 \leq h < 6$	7
$6 \leq h < 8$	5

June

Number of hours sunshine h	Number of days
$0 \leq h < 2$	5
$2 \leq h < 4$	7
$4 \leq h < 6$	10
$6 \leq h < 8$	8

- a Calculate estimates of the mean number of hours of sunshine in May and in June.
b Write down
i the class which contains the median ii the modal classes for May and June.

KEY WORDS:

- interquartile range
- frequency polygon
- lower quartile
- range
- upper quartile
- stem and leaf diagram

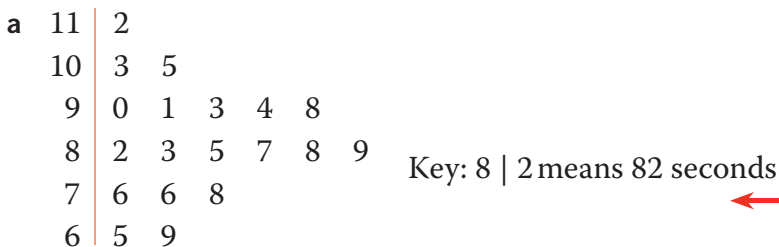
- When you use graphs to compare data sets always make comparisons between the same type of average (compare means or medians or modes) or the range or the IQR.
- The **range** of a data set = largest value – smallest value
- The **interquartile range (IQR)** = upper quartile – lower quartile (UQ – LQ)
 - To find the quartiles put the data in order, then
 - $\frac{1}{4}$ of the values \leq **LQ**
 - $\frac{3}{4}$ of the values \leq **UQ**
 - For small data sets find $\frac{1}{4}$ and $\frac{3}{4}$ of $(n + 1)$ th value.
- You can use a **stem and leaf** diagram to display small data sets.
 - The stem is written on one side of a vertical line with leaves on the other side.
 - Leaves are written in order with the smallest next to the stem.
 - Always write a key.

EXAMPLE

The times taken, in seconds, for 19 children to complete a jigsaw are

69 103 94 65 88 76 78 93 105 112
83 98 85 89 91 76 87 90 82

- a Draw a stem and leaf diagram to display these data.
- b 19 adults completed the same jigsaw. Their results are summarised as median 76 seconds, IQR 20 seconds. Compare the times of the children and the adults.



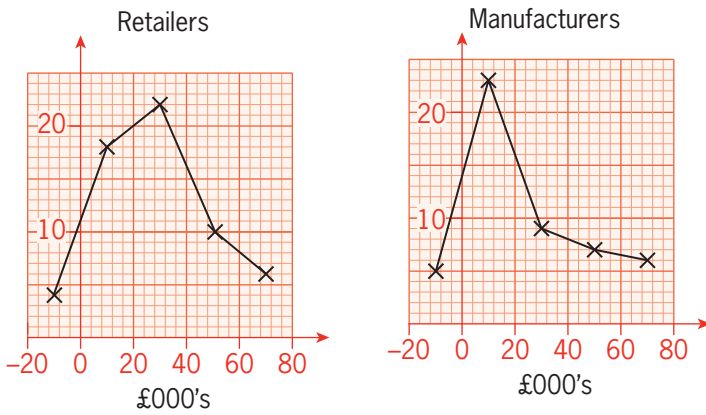
- b Median for children $\frac{1}{2}(19 + 1)$ th = 10th value = 88 seconds, so on average the adults were faster. IQR for children $94 - 78 = 16$ seconds which is lower than the adult's so the children's results were more consistent.

Use tens as the stem and units as the leaves. 112 is 11 tens and 2 units. Choose any value in the diagram for the key.

- You can use a **frequency polygon** to display continuous data that has been grouped into classes.
 - Plot each frequency with the midpoint of the class.
 - Join the points with straight lines using a ruler.

EXAMPLE

Compare the profits (or losses) for these samples of retailers and manufacturers.



Don't be tempted to compare individual values such as the last class in both graphs.

Don't be tempted to compare the height of the peak compare the classes of the highest peak from each graph

On average the retailers made more profit as their modal class (£20 000-£40 000) is higher than the modal class (£0-£20 000) for the manufacturers.

Exercise S3

- These are the times taken, to the nearest minute, to deliver leaflets to 15 roads.
22 15 28 34 29 9 11 24 23 16 20 30 31 32 41
Represent these data on a stem and leaf diagram.
- The average attendance, to the nearest 100, at premier league football games in 2009/10 were as follows.
4600 4300 4100 2800 3900 3700 6000 3600 2400 4000
1800 2500 2700 3400 2400 1800 2100 2500 7500 2200
Represent these data on a stem and leaf diagram.
- The back-to-back stem and leaf diagram shows the lengths of samples of two different species of caterpillar.

Species A		Species B
5 4	7	2
9 4 3	6	0 1 5
7 7 5	5	2 3 4 6 7 8
9 8 5 3	4	0 1 9 9
7 5 4 3	3	1 4 5 6
6 4 1	2	9

Key: 1 | 2 | 9 means 21 mm for species A and 29 mm for species B Make three comparisons between these samples of species of caterpillars.

Overmatter exercise S3

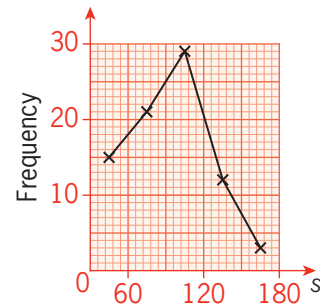
- 4 Todd measured the lengths, in mm, of leaves, dropped from trees in his garden one day in autumn. The data he collected is summarised in the table.
- Draw a frequency polygon to show the length of the leaves.
 - Explain how the data suggests that there may be more than one type of tree in Todd's garden.

Leaf length, l mm	Frequency
$20 < l \leq 30$	2
$30 < l \leq 40$	3
$40 < l \leq 50$	11
$50 < l \leq 60$	0
$60 < l \leq 70$	0
$70 < l \leq 80$	6
$80 < l \leq 90$	5
$90 < l \leq 100$	3

- 5 In a triathlon there are transition times when competitors change sport from swim to bike and then from bike to run. Paolo collected data on the transition times, in seconds, for 80 competitors. These are the transition times for swim to bike

Time, s seconds	$120 < s \leq 150$	$150 < s \leq 180$	$180 < s \leq 210$	$210 < s \leq 240$	$240 < s \leq 270$
Frequency	12	21	26	15	6

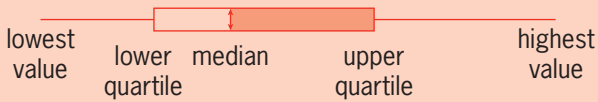
- Draw a frequency polygon to represent these data. Paolo drew this frequency polygon for the transition times for bike to run.
- Compare the transition times taken for swim to bike and for bike to run.



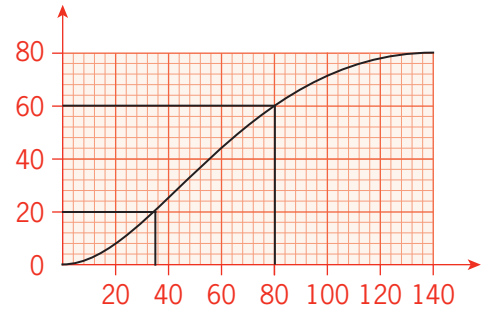
KEY WORDS:

- box plot
- cumulative frequency graph
- IQ
- IOR
- LQ
- median

You can use a **box plot** to show how data are spread



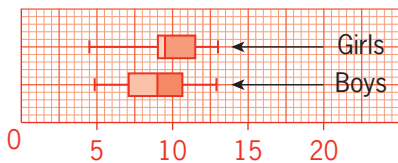
- If the **median** is closer to the **LQ** than the **UQ** the data is positively skewed
- If the median is closer to the **UQ** than the **LQ** the data is negatively skewed
- For small data sets LQ is $\frac{1}{4}$ and UQ $\frac{3}{4}$ of $(n + 1)$ th value
- For large data sets LQ is $\frac{1}{4}$ and UQ $\frac{3}{4}$ of n th value
- You use a **cumulative frequency graph** to represent large amounts of grouped data.
- You can estimate the median and the LQ, the UQ and the **IOR** from a cumulative frequency graph.
- You can use the median and IQR draw a box plot.



To find the lower quartile draw a horizontal line from $\frac{1}{4}$ of the cumulative frequency (CF) on the vertical axis across to the graph. From this point draw a vertical line down to the horizontal axis and read off the LQ = 35. Use $\frac{1}{2}$ CF to find the median = 60 and $\frac{3}{4}$ CF for the UQ = 80.

EXAMPLE

The box plots summarise the times taken by boys and girls to complete a sudoku.



Similarity: They have the same median or they have the same IQR.

Difference: Range for boys is smaller than range for girls or girls' times are negatively skewed, while boys' times are not skewed.

Give one similarity and one difference between the times taken by the girls and the times taken by the boys.

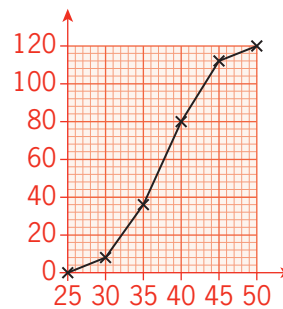
Don't be tempted to compare individual values such as only the lowest values, or different values such as girls' median greater than boys' LQ when making comparisons.

EXAMPLE

The cumulative frequency diagram summarises the times taken by 120 people to swim a mile.

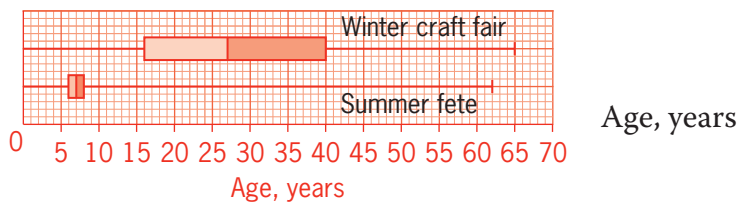
- a What was the median time?
- b This quote came from a news report. 'The range of times taken to swim a mile was 23 minutes, the quickest time was 28 minutes' Could this quote be accurate? Explain your answer

- a $37\frac{1}{2}$ minutes
- b Range 23 is possible:
 $23 < 50 - 25$
 Quickest time 28 is possible: interval 25 to 30 has 4 people
 But $28 + 23 = 51$ and 51 cannot be longest time, quote is not accurate.



Exercise S4

- 1 One year a school held a summer fete and a winter craft fair. The age distribution for each is summarised in these box plots



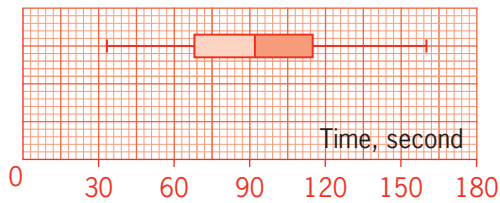
Compare the age distributions at these events.

- 2 In a triathlon there are transition times when competitors change sport from swim to bike and then from bike to run. Paolo collected data on the transition times, in seconds, for 80 competitors. These are the transition times for swim to bike.

Time, s seconds	$120 < s \leq 150$	$150 < s \leq 180$	$180 < s \leq 210$	$210 < s \leq 240$	$240 < s \leq 270$
Frequency	12	21	26	15	6

Overmatter exercise S4

- a Draw a cumulative frequency graph to represent this information.
Paolo drew this box plot for the transition times for bike to run



- b Find suitable measures from both graphs and compare the transition times taken for swim to bike and for bike to run.
- 3 The table summarises the arrival times, in minutes, of students before the start of an exam at a school.

Time, t minutes	$0 \leq t < 4$	$4 \leq t < 8$	$8 \leq t < 12$	$12 \leq t < 16$	$16 \leq t < 20$	$20 \leq t < 24$	$24 \leq t < 28$
Frequency	6	18	24	46	64	30	12

- a Draw a cumulative frequency table.
- b Draw a cumulative frequency graph.
- c Use your graph to find estimates of the number of people who arrived
- within 10 minutes before the start of the exam
 - over 22 minutes before the exam began.
- d The first student arrived for the exam 27 minutes early, the last student arrived as the exam was about to start. Estimate the other measures needed and draw a box plot for these data.
- 4 The times taken for 120 people to complete a run are summarised in this report
‘The fastest time to complete the run was 22 minutes, with the first 18 people finishing in under $\frac{1}{2}$ hour. In total 60 people finished within 40 minutes. Only 17 people took over 50 minutes and of these five people took longer than an hour with the final time recorded as 68 minutes.’
- a Draw a cumulative frequency graph to represent this information.
- b Use the information and the graph to draw a box plot.

KEY WORDS:

- | | |
|-------------------|---------------|
| bar | histogram |
| class interval | modal class |
| class width | negative skew |
| frequency | positive skew |
| frequency density | |

The highest bar might not be the same as the class with the greatest frequency as the class intervals might not be the same.

- You use a **histogram** to represent continuous data - usually the data is grouped in classes.
- Each **bar** of the histogram represents one **class**.
- The **frequency** of a class is represented by the area of the bar.
- The width of the bar is the **class interval**.
- Class intervals can be different widths.
- The height of a bar is its **frequency density** of the class.
- Frequency density is shown on the vertical axis.
- Frequency density = frequency ÷ class width**
- On a histogram the **modal class** is the highest bar.

- If the histogram bars peak in the middle with the bars on either side approximately the same heights the data is not skewed.
- If the bars peak towards the lower end of the data (on the left of the histogram) the data has **positive skew**.
- If the bars peak towards the higher end of the data (on the right of the histogram) the data has **negative skew**.

EXAMPLE

The table summarises the speeds, v mph, of 80 vehicles travelling along Berry Road.

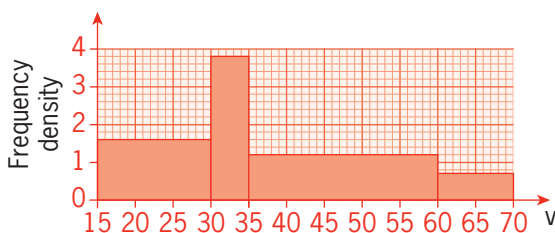
Speed v mph	$15 < v \leq 30$	$30 < v \leq 35$	$35 < v \leq 60$	$60 < v \leq 70$
Number of vehicles	24	19	30	7

- Draw a histogram to represent these data.
- The speed limit for the road is 40 mph. Estimate how many of the cars are exceeding the speed limit.
- Speeds of 80 vehicles travelling on Cherry road were found to have a median in the class $15 < v \leq 30$. Comment on the different speeds of vehicles in these two roads.

Frequency density = frequency ÷ class width

a

Class width	$30 - 15 = 15$	$35 - 30 = 5$	$60 - 35 = 25$	$70 - 60 = 10$
Frequency density	$24 \div 15 = 1.6$	$19 \div 5 = 3.8$	$30 \div 25 = 1.2$	$7 \div 10 = 0.7$



Be clear which data you are referring to when writing comments.

- Area of bars above 40 mph = $1.2 \times 20 + 0.7 \times 10 = 24 + 7 = 31$ so 31 cars were speeding.
- On Berry Road the median is in class $30 < v \leq 35$ ($\frac{1}{2}$ of 80 = 40 and $24 + 19 = 43$ 40th value must be in class $30 < v \leq 35$ compared with $15 < v \leq 30$ on Cherry Road so over half the cars are travelling faster on Berry Road than Cherry Road.

Exercise S5

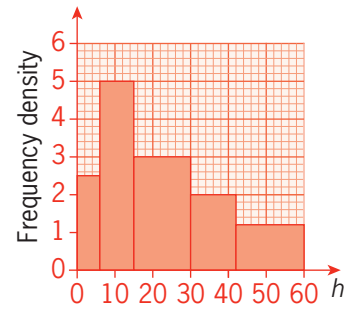
- 1 Jenny asked a group of men about their weekly consumption of units of alcohol. She summarised her results in this table.

Units of alcohol, u	0	$0 < u \leq 1$	$1 < u \leq 7$	$7 < u \leq 15$	$15 < u \leq 30$	Over 30
Number of people	8	12	27	30	21	9

- a The last group 'Over 30' is open-ended. Choose a suitable value for this last group and use it to draw a histogram to represent these data.
- b The 'low risk' consumption for men is less than 21 units of alcohol per week. Calculate an estimate of the number of men who would be considered low risk.
- 2 The table summarises survey data about the average hours men worked per week.

Hours worked, h	$0 \leq h < 6$	$6 \leq h < 15$	$15 \leq h < 30$	$30 \leq h < 42$	$42 \leq h < 60$
Frequency	3	9	27	69	42

- a Draw a histogram to represent these data.
- b Estimate the number of people who work more than 40 hours a week.
The histogram summarises survey data about the average hours women worked per week.
- c Explain how the graph shows that the same number of women worked 6 to 15 hours as 15 to 30 hours.
- d Write down two comparisons between the average hours men and women worked in the survey.



- 3 The incomplete table and incomplete diagram show the expected journey times to work for a sample of city workers.

Time, t min	$0 < t \leq 20$	$20 < t \leq 30$	$30 < t \leq 40$	$40 < t \leq 45$	$45 < t \leq 60$	$60 < t \leq 90$
Number		16		28	33	21

- a Copy and complete the table and the diagram.
- b Estimate the number of city workers in the sample whose journey time is longer than 50 minutes.

