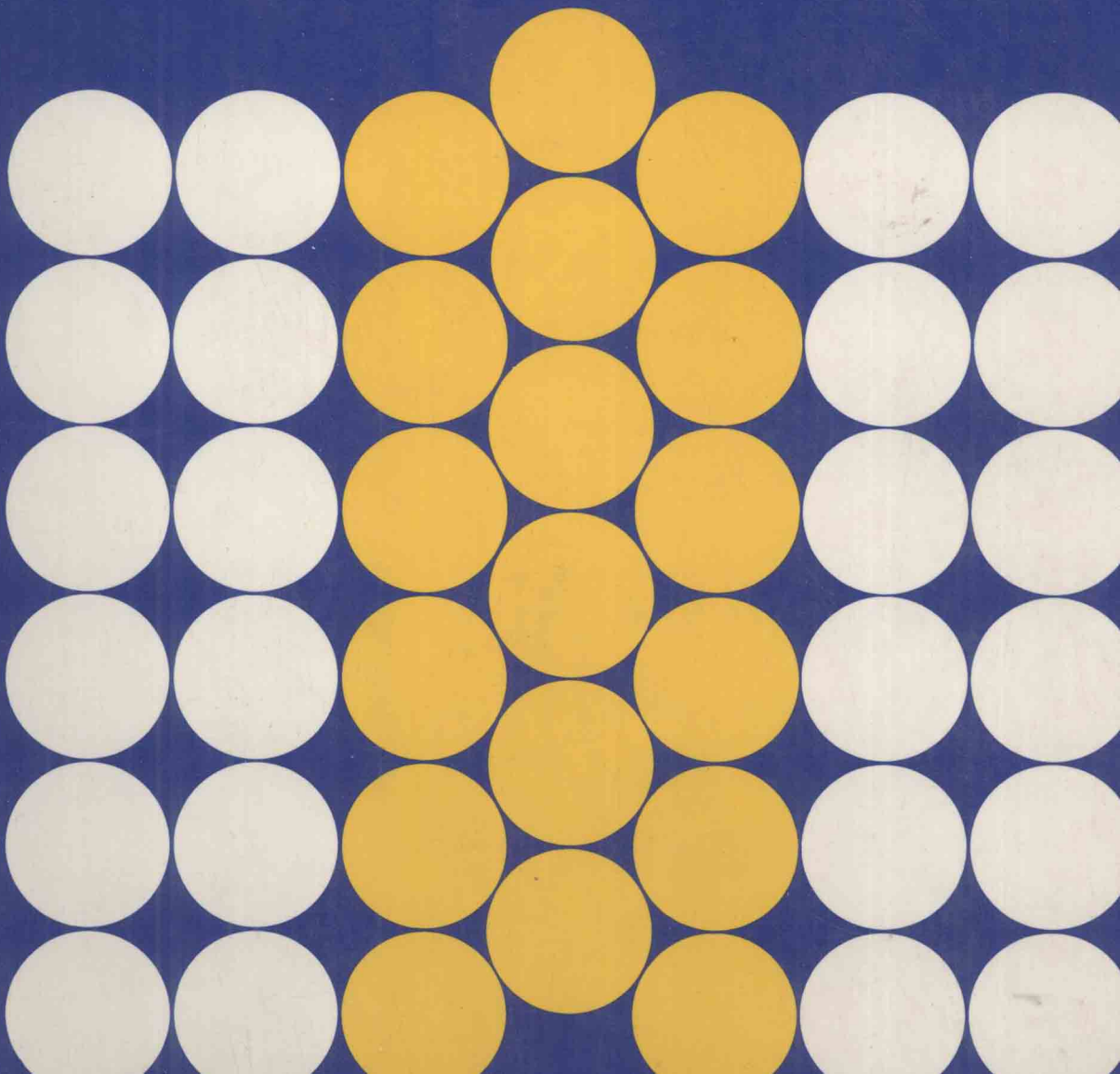
CONCISE STATISTICS

M.G. Godfrey, E.M. Roebuck and A.J. Sherlock



Concise Statistics

M G Godfrey
E M Roebuck
A J Sherlock

Edward Arnold

© M G Godfrey, E M Roebuck and A J Sherlock 1988

First published in Great Britain 1988 by
Edward Arnold (Publishers) Ltd, 41 Bedford Square, London WC1B 3DQ

Edward Arnold, 3 East Read Street, Baltimore, Maryland 21202, USA

Edward Arnold (Australia) Pty Ltd, 80 Waverley Road, Caulfield East, Victoria 3145, Australia

British Library Cataloguing in Publication Data

Godfrey, M.G.
Concise statistics.
1. Mathematical statistics—Examinations, questions, etc.
I. Title II. Roebuck, E.M. III. Sherlock, A.J.
519.5'076 QA276.2

ISBN 0-7131-3591-3

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of Edward Arnold (Publishers) Ltd.

Text set in 10/12pt Times Compugraphic by Mathematical Composition Setters Ltd, Salisbury
Printed and bound in Great Britain by J. W. Arrowsmith Ltd, Bristol.

Contents

此为试读,需要完整PDF请访问: www.ertongbook.com

1
Frequency Distributions

1.1 Histograms

A frequency distribution for a quantitative (i.e. numerical) variable can be illustrated by a histogram, which is a special type of bar chart where the *area* of the bars represents the frequency.

The variable of interest is represented on the horizontal axis, which is a continuous scale labelled in the usual manner. Bars are drawn vertically corresponding to the 'classes' into which the values have been grouped. There should be no gaps between the bars, and the precise dividing points, the *class boundaries*, are found as follows.

(a) For a continuous variable

A continuous variable is one which can take any value in a certain range, for example, lengths, weights, times and so on. For this we consider what exact values of the variable would fall into each class.

Example 1

Length (to nearest metre)	120–124	125–129	...
Frequency	8	17	

Any length between 119.5 m and 120.5 m will be recorded as 120 m (to the nearest metre), and similarly any length between 119.5 m and 124.5 m will fall into the first class. The class boundaries are

| 119.5 and 124.5 | 124.5 and 129.5 |

Note that a length of *exactly* 124.5 m is rounded up to 125 m (to the nearest metre), and therefore belongs to the second class. However any length which is just less than 124.5 m, however close (e.g. 124.499999 m) is rounded down to 124 m (to the nearest metre), and therefore belongs to the first class.

Example 2

Weight in kg (correct to 2 d.p.)	30.20–30.29	30.30–30.39
Frequency	15	27

A weight of 30.20 kg (correct to 2 d.p.) means that the exact weight is between 30.195 kg and 30.205 kg. The class boundaries are

| 30.195 and 30.295 | 30.295 and 30.395 |

Example 3

Age (in completed years)	0–5	6–12	13–18
Frequency	28	37	12

A child is said to be 5 years old when he is between his 5th and 6th birthdays, so his exact age is between 5 and 6 years.

The class boundaries are

0 and 6	6 and 13	13 and 19

(b) For a discrete variable

Now suppose that the underlying variable is *discrete* (i.e. we can list its possible values). We can draw a histogram only if we are prepared to treat the variable as a continuous one.

For example, if the possible values are the whole numbers 10, 11, 12, 13, ...
the value 11 corresponds to the interval 10.5 to 11.5
the value 12 corresponds to the interval 11.5 to 12.5 and so on.

If the possible values are (as in shoe sizes) $5, 5\frac{1}{2}, 6, 6\frac{1}{2}, \ldots$
the value $5\frac{1}{2}$ corresponds to the interval 5.25 to 5.75
the value 6 corresponds to the interval 5.75 to 6.25, and so on.

Example 4

Goals scored	0	1	2	...
Frequency	26	18	13	

The class boundaries are

−0.5 and 0.5	0.5 and 1.5	1.5 and 2.5

Example 5

Number of enquiries	0–9	10–19	20–29
Frequency	10	22	17

The class boundaries are

−0.5 and 9.5	9.5 and 19.5	19.5 and 29.5

Example 6

Shoe size	$4–5\frac{1}{2}$	$6–7\frac{1}{2}$	$8–9\frac{1}{2}$
Frequency	45	89	120

The class boundaries are

| 3.75 and 5.75 | 5.75 and 7.75 | 7.75 and 9.75 |

Note that, in all cases, the upper boundary of one class is equal to the lower boundary of the next class.

Sometimes it is necessary to make reasonable assumptions in order to determine the class boundaries.

Example 7

Height (cm)	150–	160–	170–
Frequency	4	16	24

Here '150–' clearly means the class beginning with 150; but height is a continuous variable and the accuracy to which the heights have been measured is not specified.

It could be to the nearest cm (in which case the lower class boundary is 149.5 cm)

or to the nearest 0.1 cm (in which case the lower class boundary is 149.95 cm)

and so on.

We shall assume that the heights are given exactly. The third class has no upper limit; it is reasonable to assume that it is the same size as the first two.

The class boundaries are | 150 and 160 | 160 and 170 | 170 and 180 |

Example 8

Height (cm)	150–159	160–169	170–179
Frequency	4	16	24

Again the accuracy of measurement is not specified, but since the first class ends with 159 cm and the second begins with 160 cm, it is clear that the measurements are to the nearest cm.

The class boundaries are

| 149.5 and 159.5 | 159.5 and 169.5 | 169.5 and 179.5 |

Example 9

Time (seconds) mid-interval value	35.5	55.5	75.5
Frequency	11	20	15

Here we are given the value at the centre of each class.
We assume that all the classes have the same width, and since the centres are 20 s apart, the width is 20 s. Hence each class extends 10 s each side of its centre.
The class boundaries are | 25.5 and 45.5 | 45.5 and 65.5 | 65.5 and 85.5 |

Frequency density

The **class width** is the difference between the class boundaries, and this is the width of the bar drawn on the histogram.

If the area of the bar is to be the frequency,

we have height × width = frequency, and so $\text{height} = \dfrac{\text{frequency}}{\text{width}}$

We define $$\text{frequency density} = \frac{\text{frequency}}{\text{class width}}$$

and this gives the vertical scale on the histogram.

Since the area of the histogram is equal to the frequency, we can find frequencies and probabilities as shown in Example 10.

Example 10

The heights of the 400 trees in a small copse are as follows:

Height (nearest m)	5–9	10–11	12–13	14–16	17–19	20–22	23–26	27–36
Number of trees	18	58	62	72	57	42	36	55

Draw a histogram, and use it to find
(i) the number of trees with heights between 12 m and 25 m
(ii) the probability that one of these trees, chosen at random, is taller than 25 m.

Table 1.1 shows the same data with frequency densities calculated.

Table 1.1

Height	Frequency	Class boundaries	Class width	Frequency density
5–9	18	4.5 and 9.5	5	3.6
10–11	58	9.5 and 11.5	2	29
12–13	62	11.5 and 13.5	2	31
14–16	72	13.5 and 16.5	3	24
17–19	57	16.5 and 19.5	3	19
20–22	42	19.5 and 22.5	3	14
23–26	36	22.5 and 26.5	4	9
27–36	55	26.5 and 36.5	10	5.5
	400			

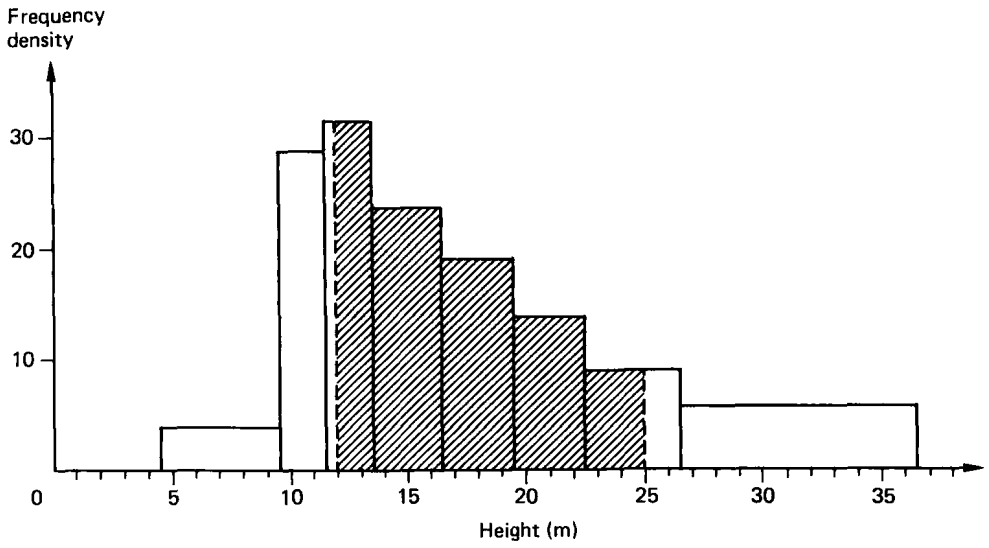


Fig. 1.1

(i) As height is a continuous variable, and the accuracy is not specified, we shall assume that 'between 12 m and 25 m' means between exactly 12 m and exactly 25 m. We require the area of the histogram between heights 12 m and 25 m, which is

$$(1.5 \times 31) + 72 + 57 + 42 + (2.5 \times 9) = 240$$

Hence there are 240 trees with heights between 12 m and 25 m.

(ii) The number of trees taller than 25 m is given by the area of the histogram to the right of 25 m, which is

$$(1.5 \times 9) + 55 = 68.5$$

There are 400 trees altogether, of which 68.5 are taller than 25 m. The probability that one tree chosen at random is taller than 25 m is therefore

$$\frac{68.5}{400} \approx 0.171.$$

Notes (1) These answers are only approximations, since we do not know how the heights are distributed within the given classes. In (ii) it is of course impossible to have 68.5 trees taller than 25 m.

(2) As the data are given to the nearest metre, we could have interpreted 'between 12 m and 25 m' to mean between 12 m and 25 m (inclusive) when measured to the nearest metre. This means that the exact height is between 11.5 m and 25.5 m, so we should find the area of the histogram between 11.5 m and 25.5 m.

The mode

The mode is the value which occurs most often; for a continuous variable it is the value with the highest frequency density.

此为试读,需要完整PDF请访问: www.ertongbook.com

For grouped data, the **modal class** is the class with the highest frequency density (i.e. having the tallest bar on the histogram). For the distribution of tree heights in Example 10, the modal class is 12–13 m.

Note that this is *not* necessarily the class with the highest frequency.

If we require a single value for the mode, we would clearly choose a value in the modal class (i.e. between 11.5 m and 13.5 m).

Since the next bar on the left is taller than that on the right, it is reasonable to choose a value closer to 11.5 m than to 13.5 m (see Fig. 1.2). Using the 'criss-cross' method on the histogram, we obtain

the mode is 11.9 m

This is, of course, only an approximate value.

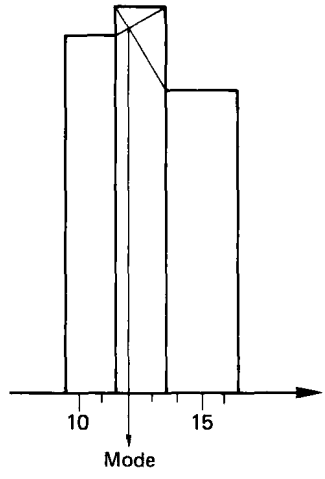


Fig. 1.2

The median

When the values are arranged in order, the median is the value in the centre. Since half the values are less than the median, and half are greater, the median can be found on a histogram by dividing the area in half.

For the distribution of tree heights in Example 10, the total area of the histogram is 400 and the area of the first three bars is $18 + 58 + 62 = 138$; we therefore need an area of 62 from the fourth bar (Fig. 1.2).

We have $24x = 62$

$$x \approx 2.6$$

and so the median is

$$13.5 + 2.6 = 16.1 \text{ m}$$

Again this is only an approximate value.

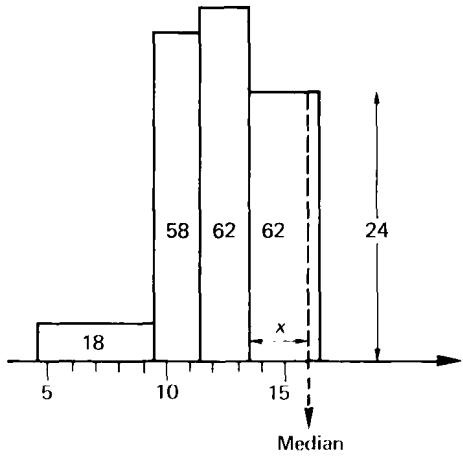


Fig. 1.3

Probability density

We define

$$\text{probability density} = \frac{\text{probability}}{\text{class width}}$$

$$= \frac{\text{frequency}}{(\text{total frequency}) \times (\text{class width})}$$

Table 1.2 shows the calculation for the distribution of tree heights in Example 10.

Table 1.2

Class boundaries	Frequency f	Class width w	Probability density $\frac{f}{400w}$
4.5 and 9.5	18	5	0.009
9.5 and 11.5	58	2	0.0725
11.5 and 13.5	62	2	0.0775
13.5 and 16.5	72	3	0.06
16.5 and 19.5	57	3	0.0475
19.5 and 22.5	42	3	0.035
22.5 and 26.5	36	4	0.0225
26.5 and 36.5	55	10	0.01375

A histogram may be drawn with probability density on the vertical scale.

Since probability density $= \dfrac{\text{frequency density}}{\text{total frequency}}$ this has exactly the same shape as the original histogram (with frequency density on the vertical scale).

The area of the histogram now gives the probability.

Since the total probability is always one, the total area of this histogram is one.

Exercise 1.1 Histograms

1 The numbers of goals scored in 42 football matches were

3, 3, 1, 2, 2, 5, 1, 2, 2, 2, 1, 4, 4, 3, 3, 3, 6, 1, 2, 2, 1,
0, 2, 7, 1, 3, 2, 3, 3, 1, 4, 3, 0, 3, 5, 2, 1, 4, 1, 3, 2, 4.

Give a frequency table and draw a histogram.

2 For each of the following, state the limits on the horizontal scale between which the histogram bars should be drawn.

(a) *Number of wickets taken*

0
1
2

(b) *Golf scores*

66–70
71–75
76–80

(c) *Marks in a test*

0–4
5
6

(d) *Shoe size*

3–5½
6
6½

(e) *Heights (nearest cm)*

150–159
160–169
170–179

(f) *Resistance (nearest 0.1 ohm)*

5.0–5.4
5.5–5.9
6.0–6.4

(g) *Weights (nearest 5 kg)*

0–25
30–50
55–75

(h) *Ages (completed yrs)*

11–13
14–16
17–19

(i) *Marks in an exam*

0–10
11–20
21–30

(j) *Price of 500 g butter (p)*

38–39
40–41
42–43

(k) *Lengths (nearest mm)*

110–
120–
130–

(l) *Reaction time in s*

0.2–
0.4–
0.6–

(m) *Diameters in cm (mid-interval value)*

14.5
24.5
34.5

(n) *Error (correct to 2 d.p.)*

$-0.10 \rightarrow -0.05$
$-0.04 \rightarrow +0.04$
$+0.05 \rightarrow +0.10$

3 Draw a histogram for the following examination results

Marks	0–20	21–30	31–40	41–45	46–50	51–55	56–60	61–70	71–100
Number of candidates	14	9	15	11	18	14	10	16	24

Use your histogram to estimate the number of candidates who scored between 43 and 54 marks (inclusive).

4 The heights of 125 children were measured to the nearest 10 cm with the following results.

Height	50–70	80–100	110–120	130–140	150–170
Number of children	18	24	23	33	27

Draw a histogram.
If one of these children is selected at random, estimate the probability that his height is between 112 cm and 128 cm (when measured exactly).

5 (i) The ages of the people living in a village are as follows

Age (in completed years)	0–9	10–19	20–34	35–54	55–79
Number of people	440	480	630	440	150

Draw a histogram with probability density on the vertical scale.
(ii) The age distribution in a second village is

Age (in completed years)	0–3	4–23	24–38	39–48	49–58	59–73	74–88
Number of people	54	180	291	315	360	384	90

Why is it difficult to compare the two villages by simply looking at these tables? Draw a histogram (with probability density on the vertical scale) for the second village, and comment on the differences between the two villages.

1.2 Cumulative frequency diagrams

A histogram gives a clear 'picture' of a distribution, but when finding frequencies, the median, and so on, it is usually easier to use a cumulative frequency diagram.

Cumulative frequency is obtained by adding together all the frequencies so far, as shown below for the distribution of tree heights in the previous section.

Height (nearest m)	5–9	10–11	12–13	14–16	17–19	20–22	23–26	27–36
Frequency	18	58	62	72	57	42	36	55
Cumulative frequency	18	76	138	210	267	309	345	400

For example, the cumulative frequency 138 is calculated as 18 + 58 + 62. This means that 138 trees have heights in the combined class 5–13, i.e. have heights less than 13.5 m. Similarly 267 trees have heights less than 19.5 m, and so on.

Thus on a cumulative frequency diagram, cumulative frequency (cf) is plotted against the **upper class boundary**. In this case we plot the points (9.5, 18), (11.5, 76), (13.5, 138) and so on. We may also plot a cumulative frequency of zero against the lower class boundary of the first class—in this case the point (4.5, 0).

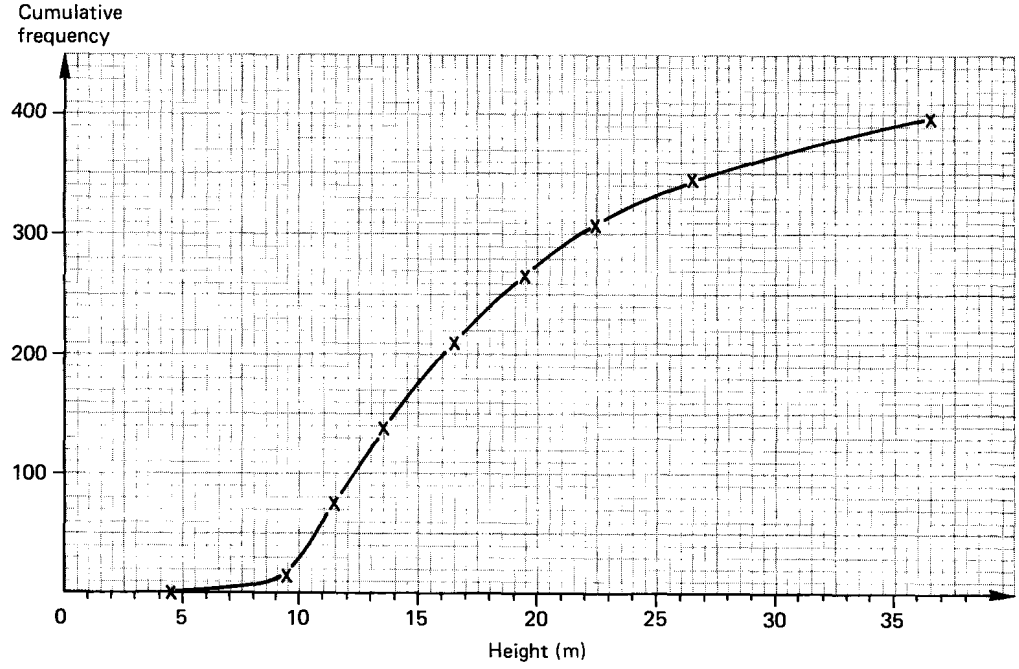


Fig. 1.4

The points may be joined by straight lines; this gives a **cumulative frequency polygon.** This assumes that the values are uniformly distributed within each class, and gives the same results as those obtained from the histogram.

However, slightly better results might be expected if the points are joined by a smooth curve; this gives a **cumulative frequency curve.**

The cumulative frequency curve for the above example is shown in Fig. 1.4.

The cumulative frequency curve can be used to find frequencies and probabilities: for any height x, the cumulative frequency gives the number of trees having height less than x.

Since half of the values are less than the median, the median can be found as the height for which the cumulative frequency is half of the total frequency.

We can also find other **percentiles**; for example the 65th percentile is the height for which the cumulative frequency is 65% of the total frequency.

Note that the median is the 50th percentile.

Example 1

For the distribution of tree heights given above, use the cumulative frequency curve to find

(i) the number of trees taller than 20 m
(ii) the probability that a tree chosen at random has a height between 15 m and 30 m
(iii) the median height
(iv) the 15th percentile.

(i) For a height of 20 m, the cumulative frequency is 275.
Thus 275 trees are shorter than 20 m, and so the number of trees taller than 20 m is $400 - 275 = 125$.

(ii) For a height 30 m, cf = 369, i.e. 369 trees are shorter than 30 m.
For a height 15 m, cf = 177, i.e. 177 trees are shorter than 15 m. So the number of trees with heights between 15 m and 30 m is $369 - 177 = 192$ and hence the probability is $192/400 = 0.48$.

(iii) For the median, $\text{cf} = \frac{1}{2} \times 400 = 200$;
hence the median is 16.0 m.

(iv) For the 15th percentile, $\text{cf} = 0.15 \times 400 = 60$;
hence the 15th percentile is 11.1 m.
This means that 15% of the trees are shorter than 11.1 m.

Linear interpolation

If we do not wish to draw an accurate cumulative frequency diagram, we can assume that the points are joined by straight lines and use similar triangles to *calculate* intermediate points.

This is *linear interpolation.*

Example 2

The marks obtained by 75 students in a test were as follows:

Mark	1–30	31–60	61–90	91–120	121–150	151–180
Number of students	3	9	20	22	13	8

Use linear interpolation to find:
(i) the median mark; (ii) the number of students who scored 140 marks or more.

It is clear that marks must be whole numbers, so
the first class has boundaries 0.5 and 30.5,
the second class has boundaries 30.5 and 60.5, and so on.
Cumulative frequency would be plotted as follows

Mark	30.5	60.5	90.5	120.5	150.5	180.5
Cumulative frequency	3	12	32	54	67	75

(i) For the median, $\text{cf} = \frac{1}{2} \times 75 = 37.5$ which comes between the points (90.5, 32) and (120.5, 54), see Fig. 1.5

We have $\frac{x}{30} = \frac{5.5}{22}$

$$x = 7.5$$

and so the median is $90.5 + 7.5 = 98.0$.

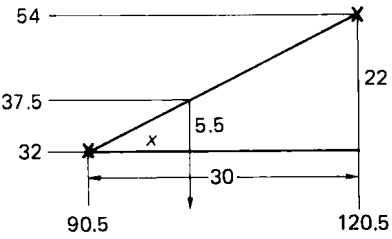


Fig. 1.5

(ii) Since marks are whole numbers, a score of 140 marks or more corresponds to 'greater than 139.5' on the continuous scale.
We consider the two points (120.5, 54) and (150.5, 67) (Fig. 1.6).

We have $\frac{y}{13} = \frac{19.0}{30.0}$

$$y \approx 8.2$$

so the cumulative frequency is

$$54 + 8.2 = 62.2.$$

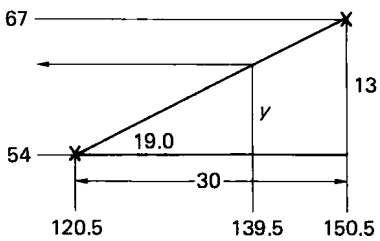


Fig. 1.6

Hence the number of students who scored 140 marks or more is $75 - 62.2 = 12.8$

These answers are only approximations, since we do not know the actual marks.

Note When the marks are arranged in order, the median is the 38th mark in the list. However, this is best estimated by interpolating to 37.5 on the cumulative frequency scale.

Cumulative probability

If we divide the cumulative frequencies by the total frequency, we obtain the **cumulative probabilities.**

For the distribution of tree heights given earlier, a cumulative probability diagram

此为试读,需要完整PDF请访问: www.ertongbook.com