Chapter 16

Statistics

Statistics involves the collection, organisation, presentation and interpretation of numerical information.

Information can be collected, and is called **data**.

A **variable** is a quantity which can vary. Variables are represented by *uppercase* letters (usually X).

The values of a **discrete** variable can be written in a list, and have gaps between them.

In contrast, a **continuous** variable can equal any real number from some particular interval, and so does **not** have gaps between its possible values.

Example 1. Suppose we survey 60 families to find out

(a) the number of children in each family, and

(b) the height of the children in each family.

If we let

$$X = \text{the number of children in a family,}$$

and

$$Y = \text{the height of a child, measured in centimetres,}$$

then X and Y are **variables**.

The variable X could equal any of the values $0,\ 1,\ 2,\ 3,\ 4,\ \ldots\ ,$ and the variable Y could equal any of the values between, say, 50 and 170.

□

Notice that the variable X defined above is **discrete**, whereas Y is **continuous**.

16.1 Tabulation and Representation of Data

Data is often organised into one of

- a **frequency** distribution
- a **relative frequency** distribution, in which frequencies are expressed as *proportions*
- a **cumulative frequency** distribution, obtained by *adding* up the frequencies.

Example 2. (Discrete Data)

Consider the variable X defined to be the number of children in a family.

Suppose that after surveying 60 families, the data shown in the adjacent table was obtained.

That is, 10 of the families contained 0 children,
4 of the families contained 1 child,
18 of the families contained 2 children, etc.

Calculate the relative frequencies and cumulative frequencies for this data.

X	Frequency
0	10
1	4
2	18
3	12
4	6
5	5
6	5
Total	60

Solution:

X	Frequency	Relative Frequency
0	10	$\frac{10}{60}$
1	4	$\frac{4}{60}$
2	18	$\frac{18}{60}$
3	12	$\frac{12}{60}$
4	6	$\frac{6}{60}$
5	5	$\frac{5}{60}$
6	5	$\frac{5}{60}$
Total	60	$\frac{60}{60} = 100\%$

X	Cumulative Frequency
< 0	0
< 1	10
< 2	14
< 3	32
< 4	44
< 5	50
< 6	55
≤ 6	60

□

Note that when we construct the intervals in a frequency distribution table,

- we usually like to have between 5 and 15 intervals
- we often prefer to make the intervals of equal width

- we should ensure that each observation fits into **exactly one** of the intervals.

Example 3. (Continuous Data)

Consider the variable Y defined to be the height of a child, measured in centimetres.

Suppose that after measuring the height of 155 children, the data shown in the adjacent table was obtained.

Calculate the relative frequencies and cumulative frequencies for this data.

Y	Frequency
50–	22
70–	38
90–	39
110–	28
130–	19
150 – 170	9
Total	155

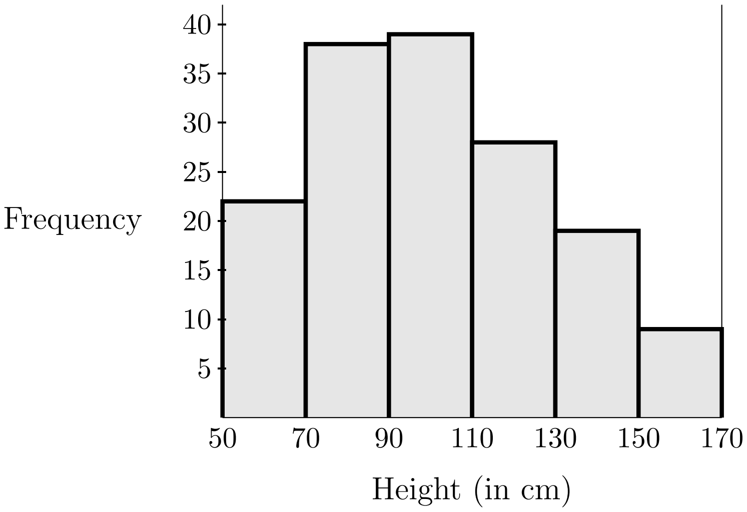
Solution:

Y	Frequency	Relative Frequency
50–	22	$\frac{22}{155}$
70–	38	$\frac{38}{155}$
90–	39	$\frac{39}{155}$
110–	28	$\frac{28}{155}$
130–	19	$\frac{19}{155}$
150 – 170	9	$\frac{9}{155}$
Total	155	$\frac{155}{155} = 100\%$

Y	Cumulative Frequency
< 50	0
< 70	22
< 90	60
< 110	99
< 130	127
< 150	146
≤ 170	155

□

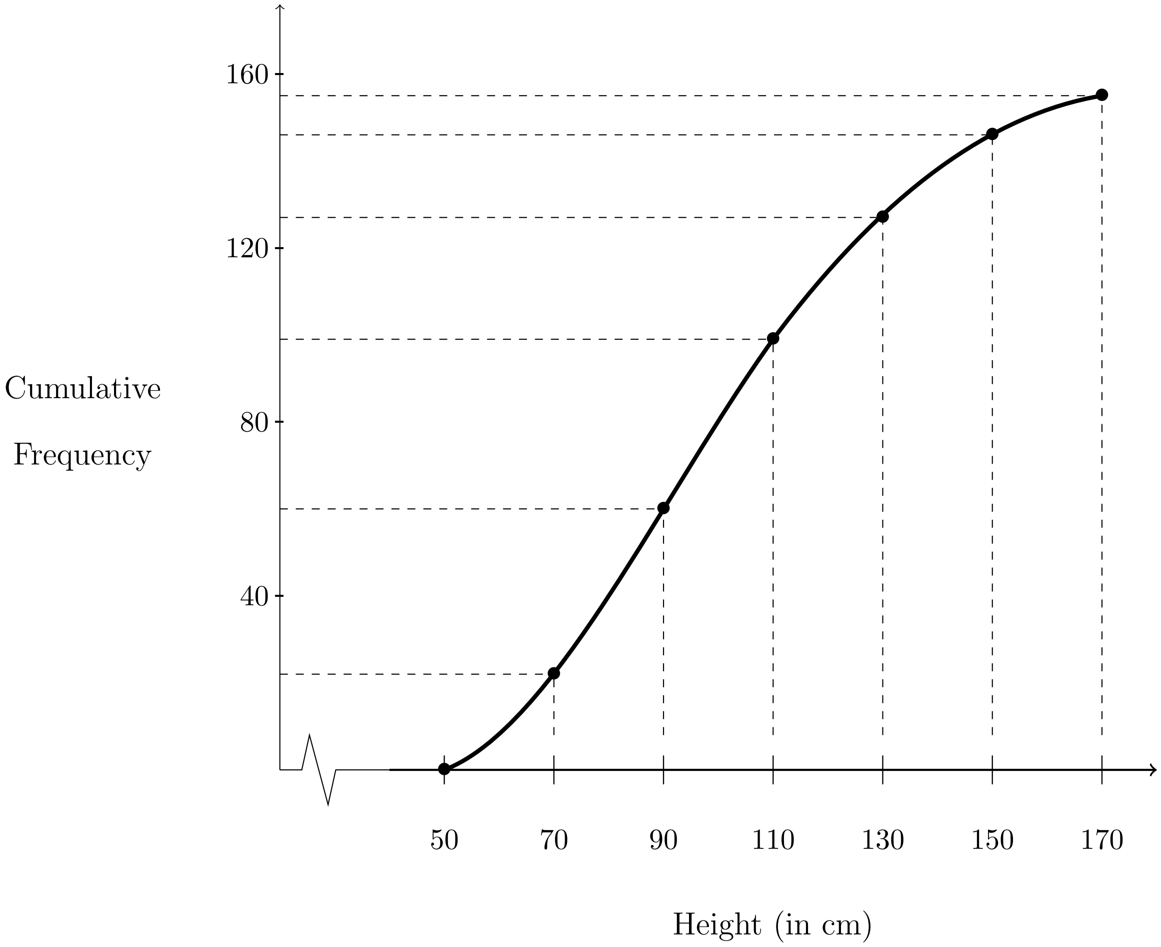
We can represent the data from this example pictorially by a **frequency histogram**, as shown:



We have already seen that the cumulative frequency distribution for the previous example was as follows:

Y	Cumulative Frequency
< 50	0
< 70	22
< 90	60
< 110	99
< 130	127
< 150	146
≤ 170	155

This information can be used to draw a **cumulative frequency curve**, as shown below:



Exercises for Section 16.1

1. The following results were obtained by 60 students sitting for a history test.

81	78	75	74	66	52	52	63	54	61	63	68	43	79	78
55	58	64	65	84	58	51	47	57	33	64	57	48	56	36
42	86	41	62	32	65	61	59	77	52	59	67	77	68	64
68	54	43	12	35	53	57	63	58	75	92	69	54	55	61

   (a) Set up a frequency distribution table for the above results, using the class intervals

$$0- \;,\; 10- \;,\; 20- \;,\; 30- \;,\; \ldots \;,\; 90 - 100.$$

   (b) Construct a histogram to represent the frequency distribution obtained in (a).

2. The rental rate per week of two-bedroom flats in Melbourne was investigated in 200 randomly selected cases during 1999.

   The adjacent frequency table describes the outcome of this investigation.

Weekly Rental (in dollars)	Number of Flats
80–	6
100–	10
120–	36
140–	62
160–	34
180–	25
200–	16
220–	8
240–	1
260 – 280	2

   (a) Draw a histogram to represent the distribution of rental rates.

   (b) Construct a cumulative frequency table for this distribution.

   (c) Plot the cumulative frequency curve, and from the curve,

      i. estimate the number of flats for which the rental was less than $170 per week;

      ii. estimate the number of flats for which the rental was at least $190 per week;

      iii. estimate the rental below which 20% of the flat-dwellers paid.

16.2 Measures of Location

The mode, the median and the mean are three statistics used to describe 'where' numerical observations 'are'.

The **mode** is the value which occurs most frequently.

The **median** is the midpoint of a distribution.

The **mean** is just another word for the average.

Example 4. Suppose an intake has 500 students and Trinity asks each student

'How many (first) cousins do you have?'.

Let X be the discrete variable for the number of cousins a particular student has. The frequency distribution of X is shown below:

value of X	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
frequency	32	9	42	50	41	30	71	82	47	39	7	27	9	0	11	3

Find the mode, median and mean number of cousins.

Solution:

We see from the table above that the **mode** is 7. That is, in the survey the most frequently occurring number of cousins is 7.

Since there are 500 observations, the midpoint is between the 250^{th} and 251^{st} observations. Counting observations from an end of the distribution, we see that both of the central observations are 6. Therefore, in the survey the **median** number of cousins is 6.

To determine the mean, we multiply each value by its frequency, then sum and divide the total by 500. We obtain

$$\frac{1}{500}\Big(0 \times 32 + 1 \times 9 + \ldots + 15 \times 3\Big) = 5.916$$

and so the **mean** number of cousins is 5.916 in this survey.

□

We will consider the mean and median in more detail.

Note: The symbol $\sum$ is used for a **sum** of values.

For example, the sum of the values x_1, x_2 and x_3 can be written as $\displaystyle\sum_{i=1}^{3} x_i$.

Mean

Suppose we have n observations, denoted by $x_1, x_2, \ldots, x_n$.
The mean of these x–values is denoted by $\overline{x}$, and is given by

$$\begin{aligned} \overline{x} &= \frac{1}{n}\Big(x_1 + x_2 + \ldots + x_n\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \ . \end{aligned}$$

When the values of a discrete variable are presented in a frequency distribution table, then the following formula is convenient for finding the mean:

$$\overline{x} = \frac{\displaystyle\sum_{i=1}^{k} x_i f_i}{\displaystyle\sum_{i=1}^{k} f_i} \ .$$

X	Frequency
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_k	f_k
Total	$\displaystyle\sum_{i=1}^{k} f_i$

This formula is written on the Mathematics 1 Formula Sheet as shown here in the box:

$$\boxed{\overline{x} = \frac{\sum xf}{\sum f}}$$

Example 5. Calculate, to one decimal place, the mean of the six values 20, 22, 30, 22, 26 and 28 .

Solution:

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{6}\,(20 + 22 + 30 + 22 + 26 + 28) \\ &= \frac{74}{3} = 24.7 \text{ (1 d.p.)} \end{aligned}$$

□

Median

Half the observed values are less than or equal to the median, and half the observed values are greater than or equal to the median.

To determine a median, it is helpful to order the observations from least to greatest.

When there is an even number of observations, any number in between the two central observations in the distribution is a median. Usually, we say that the average of the two central observations is the median.

Example 6. Determine the median of the values $20, 22, 30, 22, 26, 28$.

Solution:

Ordering these values from least to greatest, we have

$$20 \quad 22 \quad 22 \quad 26 \quad 28 \quad 30.$$

The number of values (6) is even.

The average of the 3^{rd} and 4^{th} observations is $\frac{1}{2}(22+26) = 24$.

Therefore, the median is 24.

□

Example 7. The quarterly incomes of seven teenagers are shown below. Find the mean quarterly income, and the median quarterly income.

$$\$2040,\ \ \$2330,\ \ \$22\,000,\ \ \$2470,\ \ \$2130,\ \ \$2220 \ \text{ and } \$1985.$$

Solution:

The mean is given by

$$\frac{1}{7}\Big(\$2040 + \$2330 + \$22\,000 + \$2470 + \$2130 + \$2220 + \$1985\Big) = \$5025\,.$$

Ordering the incomes from least to greatest gives

$$\$1985,\ \$2040,\ \$2130,\ \$2220,\ \$2330,\ \$2470 \ \text{ and } \ \$22\,000.$$

The median is the middle (4^{th}) value. Therefore, the median is $2220.

□

Example 7 illustrates that the mean can sometimes be distorted by an extreme value which is drastically different from the other observations. In this example we see that the mean is more than twice as large as 6 of the 7 incomes, and so is uncharacteristic of the group. In such cases, the median often is used as the statistical measure of the central location of the observations.

Exercises for Section 16.2

1. Find the mean for the variable that has the following distribution.

Value of variable	Frequency
0	36
1	38
2	13
3	7
4	5
5	1
Total	100

2. A group of 66 students studying for a professional examination were asked how many times they had attempted the examination before. The results were as below.

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

   (a) Set up a frequency distribution table for these observations.

   (b) Find the mean, median and mode for these observations.

3. Find the mean and median for each of the following sets of data.

   (a) 0, 0, 1, 1, 1, 2, 2, 4, 6

   (b) 1.76, 2.53, 2.68, 3.92, 4.61

   (c) $-0.03,\ 1.26,\ -1.65,\ 4.78,\ 2.64,\ 8.83$

16.3 Measures of Spread

The mean, median and mode are statistics which give us information about the centre or location of data. Separately, though, they tell us nothing about the variability or **spread** of the data. For example, the data sets

$$\{-20,\ -1,\ 5,\ 5,\ 36\} \quad \text{and} \quad \{4,\ 5,\ 5,\ 5.25,\ 5.75\}$$

have the same mean, median and mode, but not the same spread.

Range

The simplest way of measuring the spread of data is to use the **range**. The range is equal to

$$\boxed{\text{(the greatest value)} - \text{(the least value).}}$$

For the two data sets above, the range of the first is $36 - (-20) = 56$, whereas the range of the second is only $5.75 - 4 = 1.75$.

Percentiles, in particular quartiles

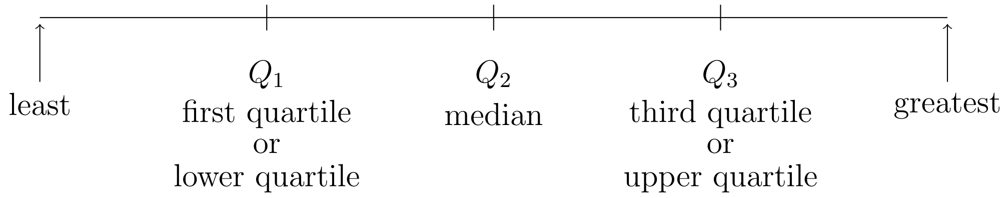
Recall that the median is the midpoint of the distribution, once the data has been ordered from least to greatest. Roughly speaking, the median breaks the data into *two* sections. Sometimes it is useful to break the distribution into more (smaller) sections.

For large data sets (such as the one about the number of cousins in Example 4), we might want to know values at particular points in the distribution. Names for those points include

- **quartiles**, at each quarter or 25% point in the distribution
- quintiles, at each fifth or 20% point
- deciles, at each tenth or 10% point
- percentiles, at each hundredth or 1% point.

In the following examples, we will consider *quartiles.* However the same method could be applied for the other points listed above.

This diagram of a distribution illustrates *quartiles*:



Alternatively, Q_1 is the median of the data from the least value up to (but not including) the median, and Q_3 is the median of the data points beyond the median up to the greatest value.

Example 8. In Example 4, we let X be the discrete variable for the number of cousins a particular student has. The frequency distribution of X is repeated below.

value of X	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
frequency	32	9	42	50	41	30	71	82	47	39	7	27	9	0	11	3

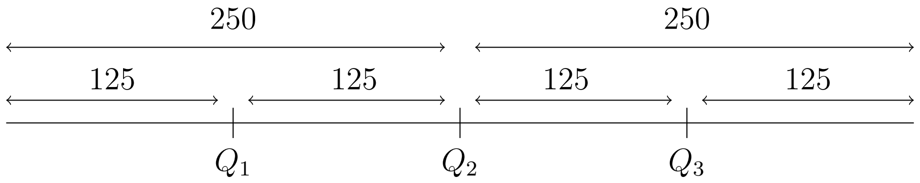
Determine the first and third quartiles for the data about cousins of 500 students.

Solution:

The median is the average of the 250^{th} and the 251^{st} observations in the ordered list. Thus there are 250 values before the median and 250 after it.

Suppose that the list has been ordered from least to greatest.

The first quartile is the median of the first 250 terms in that ordered list, and the third quartile is the median of the last 250 terms in that ordered list. Thus the first quartile is the average of the 125^{th} and 126^{th} terms from the *start* of that ordered list, and the third quartile is the average of the 125^{th} and 126^{th} terms from the *end* of that ordered list.



For the first quartile: From the frequency distribution, we see that there are 83 data points for which $X \leq 2$ and 50 for which $X = 3$. Therefore, the 125^{th} and 126^{th} data points from the *start* of the ordered list are both 3, and so the first quartile is 3.

For the third quartile, it is quickest to count back from the greatest value. From the frequency distribution, we see that there are 96 data points with $X \geq 9$ and 47 with $X = 8$. Therefore, the 125^{th} and 126^{th} data points from the *end* of the ordered list are both 8, and so the third quartile is 8.

□

The diagram in the solution above again illustrates that

the lower quartile Q_1 is the median of the observations before the median,

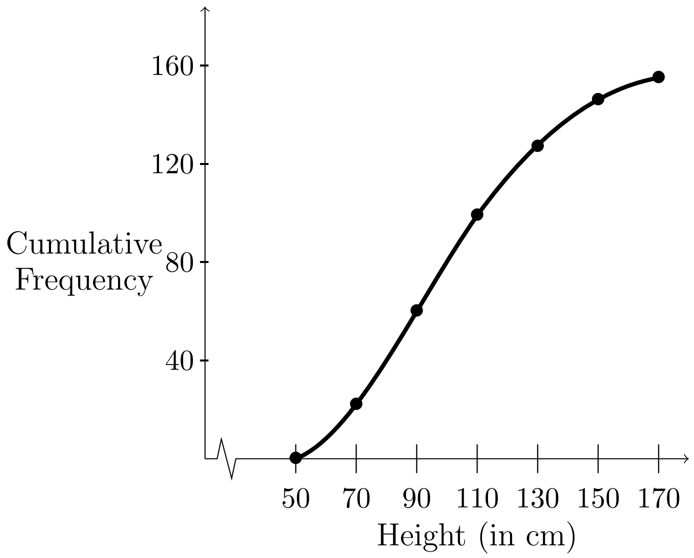
and

the upper quartile Q_3 is the median of the observations beyond the median,

in the list ordered from least to greatest.

Example 9. Consider the cumulative frequency table and the cumulative frequency curve from Example 2 in Section 16.1, as shown again here:

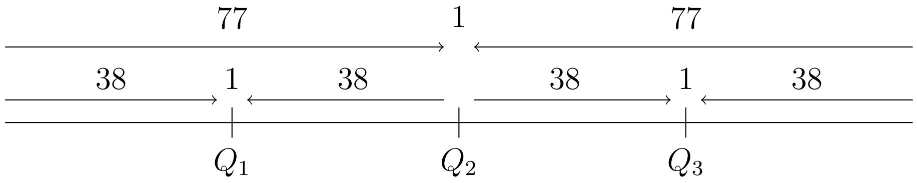
Y	Cumulative Frequency
< 50	0
< 70	22
< 90	60
< 110	99
< 130	127
< 150	146
≤ 170	155



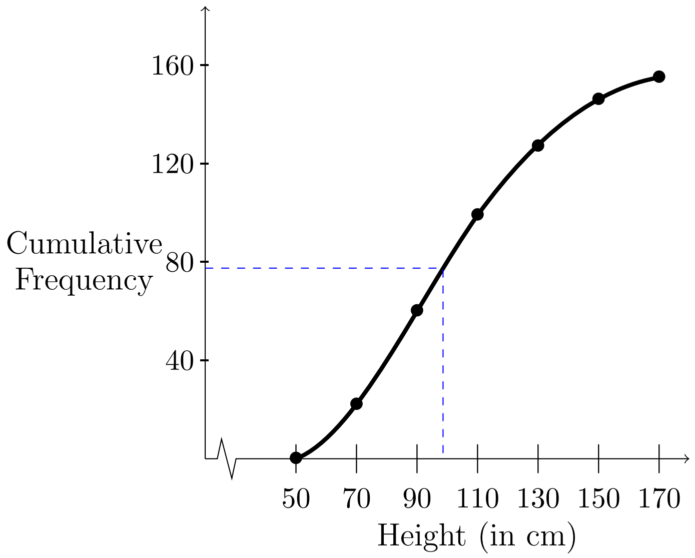
Estimate the median, and the lower and upper quartiles for this data.

Solution:

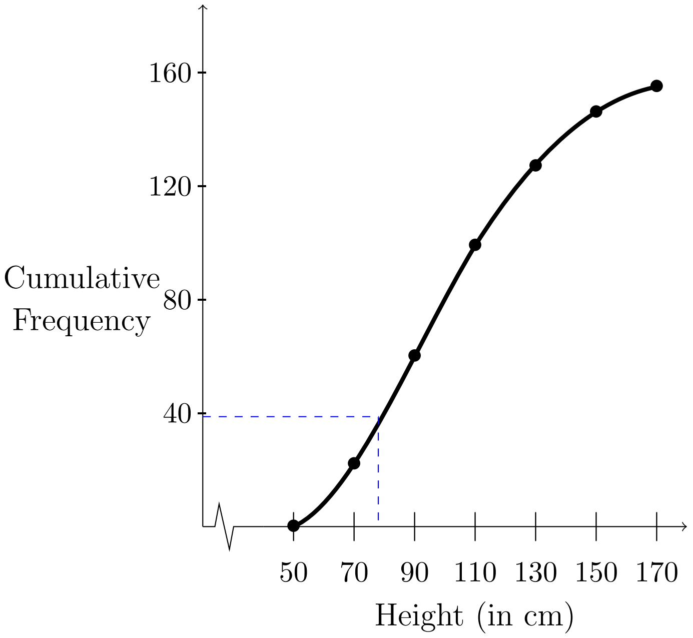
We may use a diagram to determine where to look for the quartiles amongst the **155 values** of Y.



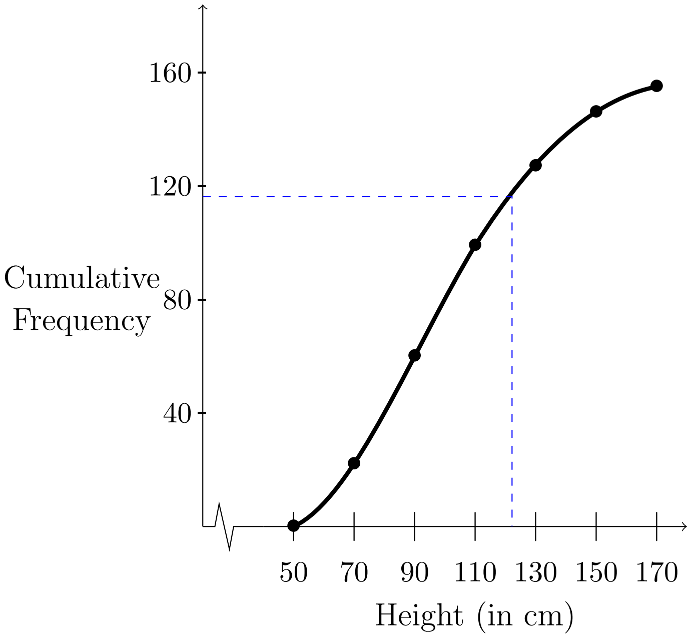
Using the cumulative frequency curve, we *estimate* the median, as the 78^{th} (ordered) observation, to be 98 cm.



Similarly, using the cumulative frequency curve, we *estimate* the lower quartile, as the 39^{th} observation, to be 78 cm.



Finally, and again using the cumulative frequency curve, we *estimate* the upper quartile, as the $(78 + 39 = 117)^{\text{th}}$ observation, to be 122 cm.



□

Interquartile range

The **interquartile range** is $Q_3 \; - \; Q_1,$ and this is another measure of the spread of the data.

Example 10. The following observations have already been arranged from least to greatest. Find the first quartile, the third quartile, and the interquartile range of the observations

$$21, \;\; 23, \;\; 24, \;\; 25, \;\; 28, \;\; 31.$$

Solution:

The median is 24.5 .

The first quartile Q_1 is the median of $21, 23$ and 24 , and so $Q_1 = 23$.

The third quartile Q_3 is the median of $25, 28$ and $31,$ and so $Q_3 = 28$.

The interquartile range is $Q_3 - Q_1 = 28 - 23 = 5$.

□

Example 11. Find the lower quartile, the upper quartile, and then the interquartile range of the observations

$$21, \;\; 23, \;\; 24, \;\; 25, \;\; 28.$$

Solution:

The median is 24 .

The lower quartile Q_1 is the median of 21 and 23. Thus $Q_1 = \dfrac{21+23}{2} = 22$.

The upper quartile Q_3 is the median of 25 and 28. Thus $Q_3 = \dfrac{25+28}{2} = 26.5$.

The interquartile range is $Q_3 - Q_1 = 26.5 - 22 = 4.5$.

□

Standard Deviation and Variance (for a sample)

Another method for measuring the spread of data is to consider the spread from the mean. The **standard deviation**, s, is approximately the average of the distances from the mean for the set of observations. The standard deviation of the n observations $x_1,\ x_2,\ \ldots,\ x_n$ is given by the formula below.

$$\boxed{\quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}\quad}$$

Data which is more "spread out" has a larger standard deviation than data which is less "spread out".

The **variance** is equal to s^2 . That is,

$$\text{variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 .$$

We can express the variance in a more convenient form by rearranging this formula:

$$\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2-2x_i\overline{x}+\overline{x}^2\right) \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+n\overline{x}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-2\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\sum x_i+n\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-\frac{2}{n}\left(\sum_{i=1}^{n}x_i\right)^2+n\frac{1}{n^2}\left(\sum_{i=1}^{n}x_i\right)^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-\frac{2}{n}\left(\sum_{i=1}^{n}x_i\right)^2+\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right] \\
&= \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right].
\end{aligned}$$

Both formulae for variance are on the Formula Sheet provided in the Mathematics 1 exams.

$$\boxed{s^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 \;=\; \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right].}$$

Example 12. The prices of a particular Apple iPod at five shops in **A**delaide are

$110, $121, $114, $123 and $117.

In five shops in **B**risbane, the same iPod costs

$59, $45, $195, $169 and $117.

It can be checked that the average, and the median, of the five ***A****delaide iPod prices is $117. Furthermore, the average, and the median, of the five* ***B****risbane iPod prices is also $117. However, we should notice that the spread of the data obtained from the* ***A****delaide shops is quite different from the spread of the data obtained from the* ***B****risbane shops.*

Find, to three decimal places,

(a) the standard deviation of the Adelaide Apple iPod prices

(b) the standard deviation of the Brisbane Apple iPod prices.

Solution:

(a) We have $n = 5$.

We can find $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$ by making a table.

x	x^2
110	12 100
121	14 641
117	13 689
123	15 129
114	12 996
585 $=\sum_{i=1}^{5} x_i$	68 555 $=\sum_{i=1}^{5} x_i^2$

$$\begin{aligned}
\text{Then } s^2 &= \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right]\\
&= \frac{1}{5\times 4}\left[5\times 68\,555 \;-\; 585^2\right]\\
&= \frac{550}{20}\ .\\
\text{That is, } s^2 &= 27.5\\
\text{and so } s &= \sqrt{27.5} = 5.244 \quad \text{(3 d.p.).}
\end{aligned}$$

(b) Again, we have $n = 5$.

We can find $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$ by making a table.

x	x^2
59	3 481
45	2 025
117	13 689
169	28 561
195	38 025
585 $= \sum_{i=1}^{5} x_i$	85 781 $= \sum_{i=1}^{5} x_i^2$

$$\begin{aligned}
\text{Then} \quad s^2 &= \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] \\
&= \frac{1}{5 \times 4}\left[5 \times 85\,781 - 585^2\right] \\
&= \frac{86\,680}{20}\ . \\
\text{That is,} \quad s^2 &= 4334 \\
\text{and so} \quad s &= \sqrt{4334} = 65.833 \quad \text{(3 d.p.)}.
\end{aligned}$$

□

We see from this example that data which is more “spread out” has a larger standard deviation than data which is less “spread out”.

Many calculators give you s (and $\overline{x}$) if you type in the observations.

Exercises for Section 16.3

1. Compute the range, the standard deviation and the variance of these samples of daily minimum temperatures (in °C) recorded during winter in three different cities. Give your answers for the standard deviations and variances to 1 decmial place.

   (a) 10, 6, 14, 18, 12 and 17

   (b) -12, -13, 2, -5 and -3

   (c) 0.4, 1.3, -0.2, 0.8, 2.1, 1.2 and 0.4

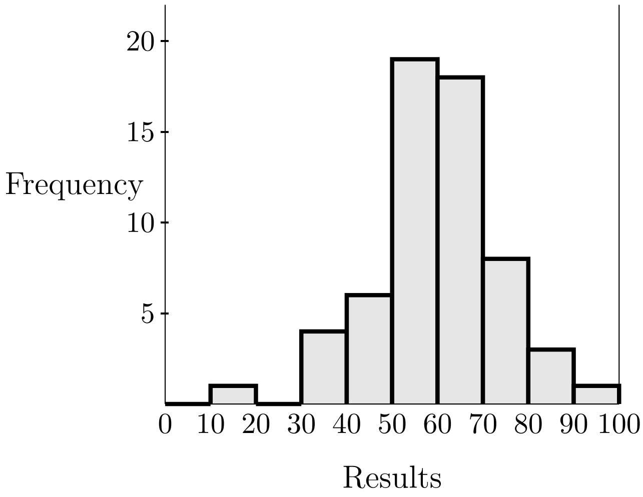
2. Find the median, the lower quartile and the upper quartile for each of the data sets given in Exercise 1 above.

16.4 Answers for the Chapter 16 Exercises

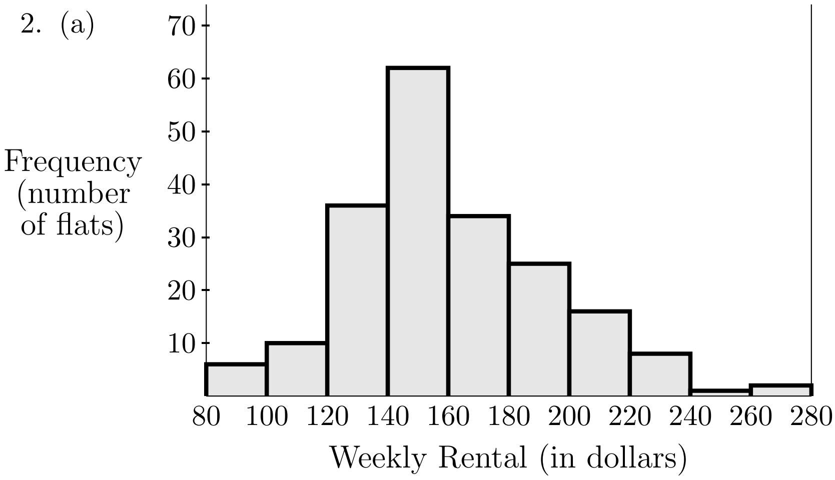
16.1 1. (a)

Results	Frequency
0–	0
10–	1
20–	0
30–	4
40–	6
50–	19
60–	18
70–	8
80–	3
90 – 100	1

(b)



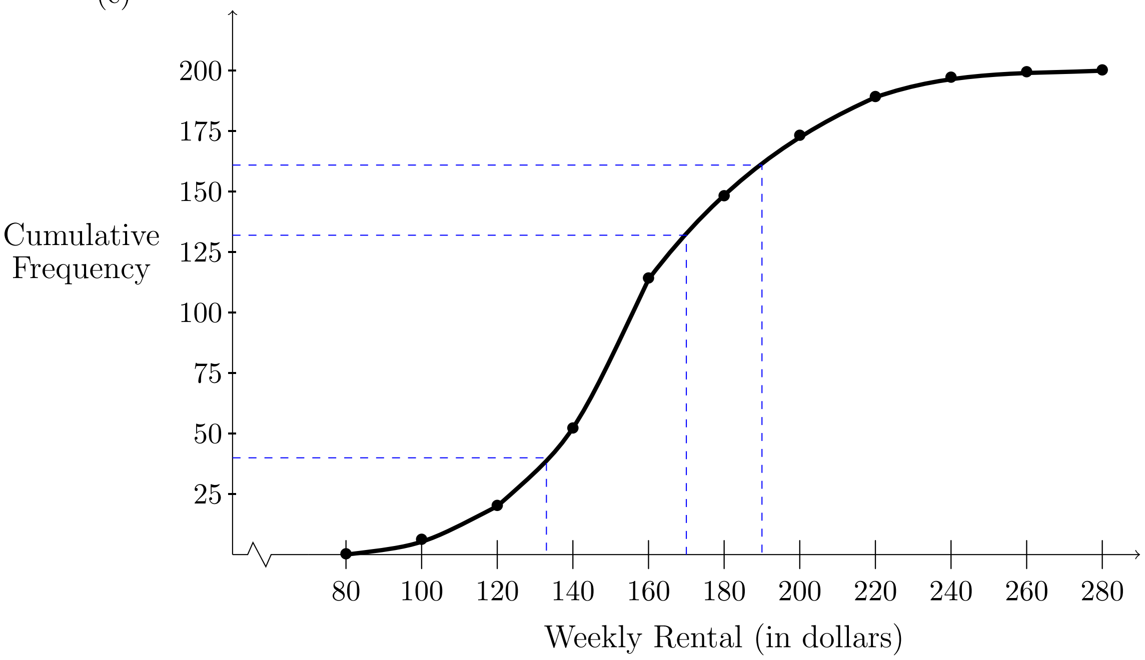
2. (a)



(b)

Weekly Rental (in dollars)	Cumulative Frequency
< 80	0
< 100	6
< 120	16
< 140	52
< 160	114
< 180	148
< 200	173
< 220	189
< 240	197
< 260	198
≤ 280	200

(c)



(i) Approximately 132 of the flats have rent less than \$170 per week.

(ii) Approximately 39 of the flats have rent of \$190 or more per week.

(iii) Twenty percent of the flat-dwellers pay less than approximately \$133 per week for their rental.

16.2 1. 1.1

2. (a)

X	Frequency
0	24
1	20
2	12
3	6
4	2
5	2

(b) mean $= \frac{40}{33}$ median $= 1$ mode $= 0$

3. (a) mean $= \frac{17}{9}$ median $= 1$
(b) mean $= 3.1$ median $= 2.68$
(c) mean $= 2\frac{383}{600}$ median $= 1.95$

16.3 1. (a) range $=$ 12 °C $s = 4.5\,°\text{C}$ (1 d.p.) $s^2 = 20.2, (°\text{C})^2$ (1 d.p.)
(b) range $=$ 15 °C $s = 6.3\,°\text{C}$ (1 d.p.) $s^2 = 39.7\,(°\text{C})^2$
(c) range $=$ 2.3 °C $s = 0.8\,°\text{C}$ (1 d.p.) $s^2 = 0.6\,(°\text{C})^2$ (1 d.p.)

2. (a) median $=$ 13 °C $Q_1 = 10\,°\text{C}$ $Q_3 = 17\,°\text{C}$
(b) median $=$ −5 °C $Q_1 = -12.5\,°\text{C}$ $Q_3 = -\frac{1}{2}\,°\text{C}$
(c) median $=$ 0.8 °C $Q_1 = 0.4\,°\text{C}$ $Q_3 = 1.3\,°\text{C}$