CHAPTER 10

Statistics

MathU Teaching® Emporium

CONTENTS

August 26, 2021

When running an experiment or conducting a survey we can potentially end up with many hundreds, thousands or even millions of values in the resulting data set. Too much data can be overwhelming and we need to reduce them or represent them in a way that is easier to understand and communicate.

Statistics is about summarising data. The methods of statistics allow us to represent the essential information in a data set while disregarding the unimportant information. We have to be careful to make sure that we do not accidentally throw away some of the important aspects of a data set.

By applying statistics properly we can highlight the important aspects of data and make the data easier to interpret. By applying statistics poorly or dishonestly we can also hide important information and let people draw the wrong conclusions.

In this chapter we will look at a few numerical and graphical ways in which data sets can be represented, to make them easier to interpret.

Statistics is used by various websites to show users who is viewing their content

1 COLLECTING DATA

DEFINITION

Data

Data refers to the pieces of information that have been observed and recorded, from an experiment or a survey.

NOTE

The word **data** is the plural of the word **datum**, and therefore one should say, "the data are" and not "the data is".

We distinguish between two main types of data: quantitative and qualitative.

DEFINITION

Quantitative data

Quantitative data are data that can be written as numbers.

Quantitative data can be discrete or continuous. Discrete quantitative data can be represented by integers and usually occur when we count things, for example, the number of learners in a class, the number of molecules in a chemical solution, or the number of SMS messages sent in one day.

Continuous quantitative data can be represented by real numbers, for example, the height or mass of a person, the distance travelled by a car, or the duration of a phone call.

DEFINITION

Qualitative data

Qualitative data are data that cannot be written as numbers.

Two common types of qualitative data are categorical and anecdotal data. Categorical data can come from one of a limited number of possibilities, for example, your favourite cooldrink, the colour of your cell phone, or the language that you learnt to speak at home

Anecdotal data take the form of an interview or a story, for example, when you ask someone what their personal experience was when using a product, or what they think of someone else's behaviour.

Categorical qualitative data are sometimes turned into quantitative data by counting the number of times that each category appears. For example, in a class with 30 learners, we ask everyone what the colours of their cell phones are and get the following responses:

black	black	black	white	purple	red	red	black	black	black
white	white	black	black	black	black	purple	black	black	white
purple	black	red	red	white	black	orange	orange	black	white

This is a categorical qualitative data set since each of the responses comes from one of a small number of possible colours.

We can represent exactly the same data in a different way, by counting how many times each colour appears.

Colour	**Count**
black	15
white	6
red	4
purple	3
orange	2

This is a discrete quantitative data set since each count is an integer.

WORKED EXAMPLE 1: QUALITATIVE AND QUANTITATIVE DATA

QUESTION

Thembisile is interested in becoming an airtime reseller to his classmates. He would like to know how much business he can expect from them. He asked each of his 20 classmates how many SMS messages they sent during the previous day. The results were

20	3	0	14	30	9	11	13	13	15
9	13	16	12	13	7	17	14	9	13

Is this data set qualitative or quantitative? Explain your answer.

SOLUTION

The number of SMS messages is a count represented by an integer, which means that it is quantitative and discrete.

WORKED EXAMPLE 2: QUALITATIVE AND QUANTITATIVE DATA

QUESTION

Thembisile would like to know who the most popular cellular provider is among learners in his school. This time Thembisile randomly selects 20 learners from the entire school and asks them which cellular provider they currently use. The results were:

Cell C	Vodacom	Vodacom	MTN	Vodacom
MTN	MTN	Virgin Mobile	Cell C	8-ta
Vodacom	MTN	Vodacom	Vodacom	MTN
Vodacom	Vodacom	Vodacom	Virgin Mobile	MTN

Is this data set qualitative or quantitative? Explain your answer.

SOLUTION

Since each response is not a number, but one of a small number of possibilities, these are categorical qualitative data.

2 MEASURES OF CENTRAL TENDENCY

2.1 Mean

DEFINITION

Mean

The mean is the sum of a set of values, divided by the number of values in the set. The notation for the mean of a set of values is a horizontal bar over the variable used to represent the set, for example $\bar{x}$. The formula for the mean of a data set $\{x_1; x_2; ...; x_n\}$ is:

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{x_1 + x_2 + \ldots + x_n}{n}\end{aligned}$$

The mean is sometimes also called the average or the arithmetic mean.

WORKED EXAMPLE 3: CALCULATING THE MEAN

QUESTION

What is the mean of the data set $\{10; 20; 30; 40; 50\}$?

SOLUTION

Step 1: Calculate the sum of the data

$$10 + 20 + 30 + 40 + 50 = 150$$

Step 2: Divide by the number of values in the data set to get the mean

Since there are 5 values in the data set, the mean is:

$$\text{mean} = \frac{150}{5} = 30$$

2.2 Median

DEFINITION

Median

The median of a data set is the value in the central position, when the data set has been arranged from the lowest to the highest value.

Note that exactly half of the values in the data set are less than the median and the other half are greater. To calculate the median of a data set, first sort the data from the smallest to the largest and then find the value in the middle. If there is an odd number of values, the median will be equal to one of the values in the data set. If there is an even number of values, the median will be halfway between two values in the data set.

WORKED EXAMPLE 4: MEDIAN FOR AN ODD NUMBER OF VALUES

QUESTION

What is the median of $\{10; 14; 86; 2; 68; 99; 1\}$?

SOLUTION

Step 1: Sort the values

The values in the data set, arranged from the smallest to the largest, are

$$1; 2; 10; 14; 68; 86; 99$$

Step 2: Find the number in the middle

There are 7 values in the data set. Since there are an odd number of values, the median will be equal to the value in the middle, namely, in the fourth position. Therefore the median of the data set is 14.

WORKED EXAMPLE 5: MEDIAN FOR AN EVEN NUMBER OF VALUES

QUESTION

What is the median of $\{11; 10; 14; 86; 2; 68; 99; 1\}$?

SOLUTION

Step 1: Sort the values

The values in the data set, arranged from the smallest to the largest, are

$$1; 2; 10; 11; 14; 68; 86; 99$$

Step 2: Find the number in the middle

There are 8 values, so there are an even number of values. The median will be halfway between the two middle values, namely, between the fourth and fifth positions, 11 and 14 respectively.

$$\text{median} = \frac{11 + 14}{2} = 12,5$$

2.3 Mode

DEFINITION

Mode

The mode of a data set is the value that occurs most often in the set. The mode can also be described as the most frequent or most common value in the data set.

To calculate the mode, we simply count the number of times that each value appears in the data set and then find the value that appears most often.

A data set can have more than one mode if there is more than one value with the highest count. For example, both 2 and 3 are modes in the data set $\{1; 2; 2; 3; 3\}$. If all points in a data set occur with equal frequency, it is equally accurate to describe the data set as having many modes or no mode.

WORKED EXAMPLE 6: FINDING THE MODE

QUESTION

Find the mode of the data set $\{2; 2; 3; 4; 4; 4; 6; 6; 7; 8; 8; 10; 10\}$.

SOLUTION

Step 1: Count the number of times that each value appears in the data set

Value	Count
2	2
3	1
4	3
6	2
7	1
8	2
10	2

Step 2: Find the value that appears most often

From the table above we can see that 4 is the only value that appears 3 times. All the other values appear less than 3 times. Therefore the mode of the data set is 4.

One problem with using the mode as a measure of central tendency is that we can usually not compute the mode of a continuous data set. Since continuous values can lie anywhere on the real line, any particular value will almost never repeat. This means that the frequency of each value in the data set will be 1 and that there will be no mode. We will look at one way of addressing this problem in the section on grouping data.

WORKED EXAMPLE 7: COMPARISON OF MEASURES OF CENTRAL TENDENCY

QUESTION

There are regulations in South Africa related to bread production to protect consumers. By law, if a loaf of bread is not labelled, it must weigh 800 g, with the leeway of 5 percent under or 10 percent over. Vishnu is interested in how a well-known, national retailer measures up to this standard. He visited his local branch of the supplier and recorded the masses of 10 different loaves of bread for one week. The results, in grams, are given below

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
$802,4$	$787,8$	$815,7$	$807,4$	$801,5$	$786,6$	$799,0$
$796,8$	$798,9$	$809,7$	$798,7$	$818,3$	$789,1$	$806,0$
$802,5$	$793,6$	$785,4$	$809,3$	$787,7$	$801,5$	$799,4$
$819,6$	$812,6$	$809,1$	$791,1$	$805,3$	$817,8$	$801,0$
$801,2$	$795,9$	$795,2$	$820,4$	$806,6$	$819,5$	$796,7$
$789,0$	$796,3$	$787,9$	$799,8$	$789,5$	$802,1$	$802,2$
$789,0$	$797,7$	$776,7$	$790,7$	$803,2$	$801,2$	$807,3$
$808,8$	$780,4$	$812,6$	$801,8$	$784,7$	$792,2$	$809,8$
$802,4$	$790,8$	$792,4$	$789,2$	$815,6$	$799,4$	$791,2$
$796,2$	$817,6$	$799,1$	$826,0$	$807,9$	$806,7$	$780,2$

WORKED EXAMPLE 7 (CONTINUED): COMPARISON OF MEASURES OF CENTRAL TENDENCY

1. Is this data set qualitative or quantitative? Explain your answer.
2. Determine the mean, median and mode of the mass of a loaf of bread for each day of the week. Give your answer correct to 1 decimal place.
3. Based on the data, do you think that this supplier is providing bread within the South African regulations?

SOLUTION

Step 1: Qualitative or quantitative?

Since each mass can be represented by a number, the data set is quantitative. Furthermore, since a mass can be any real number, the data are continuous.

Step 2: Calculate the mean

In each column (for each day of the week), we add up the measurements and divide by the number of measurements, 10.

For Monday, the sum of the measured values is $8007,9$ and so the mean for Monday is

$$\frac{8007,9}{10} = 800,8g$$

In the same way, we can compute the mean for each day of the week. See the table below for the results.

Step 3: Calculate the mean

In each column we sort the numbers from lowest to highest and find the value in the middle. Since there are an even number of measurements (10), the median is halfway between the two numbers in the middle.

For Monday, the sorted list of numbers is

$$789{,}0; 789{,}0; 796{,}2; 796{,}7; 801{,}2;$$
$$802{,}3; 802{,}3; 802{,}5; 808{,}7; 819{,}6$$

The two numbers in the middle are $801,2$ and $802,3$ and so the median is

$$\frac{801,2 + 802,3}{2} = 801,8g$$

WORKED EXAMPLE 7 (CONTINUED): COMPARISON OF MEASURES OF CENTRAL TENDENCY

In the same way, we can compute the median for each day of the week:

Day	Mean	Median
Monday	$800,8g$	$801,8g$
Tuesday	$797,2g$	$796,1g$
Wednesday	$798,4g$	$797,2g$
Thursday	$803,4g$	$800,8g$
Friday	$802,0g$	$804,3g$
Saturday	$801,6g$	$801,4g$
Sunday	$799,3g$	$800,2g$

From the above calculations we can see that the means and medians are close to one another, but not quite equal. In the next worked example we will see that the mean and median are not always close to each other.

Step 4: Determine the mode

Since the data are continuous we cannot compute the mode. In the next section we will see how we can group data in order to make it possible to compute an approximation for the mode.

Step 5: Conclusion: Is the supplier reliable?

From the question, the requirements are that the mass of a loaf of bread be between $800g$ minus 5%, which is $760g$, and plus 10%, which is $880g$. Since every one of the measurements made by Vishnu lies within this range and since the means and medians are all close to $800g$, we can conclude that the supplier is reliable.

DEFINITION

Outlier

An outlier is a value in the data set that is not typical of the rest of the set. It is usually a value that is much greater or much less than all the other values in the data set.

WORKED EXAMPLE 8: EFFECT OF OUTLIERS ON MEAN AND MEDIAN

QUESTION

The heights of 10 learners are measured in centimetres to obtain the following data set:

$$\{150; 172; 153; 156; 146; 157; 157; 143; 168; 157\}$$

Afterwards, we include one more learner in the group, who is exceptionally tall at 181 cm. Compare the mean and median of the heights of the learners before and after the eleventh learner was included.

SOLUTION

Step 1: Calculate the mean of the first 10 learners

$$\begin{aligned}\text{mean} &= \frac{150 + 172 + 153 + 156 + 146 + 157 + 157 + 143 + 168 + 157}{10} \\ &= 155,9 \text{ cm}\end{aligned}$$

Step 2: Calculate the mean of all 11 learners

$$\begin{aligned}\text{mean} &= \frac{150 + 172 + 153 + 156 + 146 + 157 + 157 + 143 + 168 + 157 + 181}{11} \\ &= 158,2 \text{ cm}\end{aligned}$$

From this we see that the average height changes by $158,2 - 155,9 = 2,3$ cm when we introduce the outlier value (the tall person) to the data set.

Step 3: Calculate the median of the first 10 learners

To find the median, we need to sort the data set:

$$\{143; 146; 150; 153; 156; 157; 157; 157; 168; 172\}$$

Since there are an even number of values, 10, the median lies halfway between the fifth and sixth values:

$$\text{median} = \frac{156 + 157}{2} = 156,5 \text{ cm}$$

Step 4: Calculate the median of all 11 learners

After adding the tall learner, the sorted data set is

$$\{143; 146; 150; 153; 156; 157; 157; 157; 168; 172; 181\}$$

Now, with 11 values, the median is the sixth value: 157 cm. So, the median changes by only $0,5$ cm when we add the outlier value to the data set.

In general, the median is less affected by the addition of outliers to a data set than the mean is. This is important because it is quite common that outliers are measured during an experiment, because of problems with the equipment or unexpected interference.

3 GROUPING DATA

A common way of handling continuous quantitative data is to subdivide the full range of values into a few sub-ranges. By assigning each continuous value to the sub-range or class within which it falls, the data set changes from continuous to discrete.
Grouping is done by defining a set of ranges and then counting how many of the data fall inside each range. The sub-ranges must not overlap and must cover the entire range of the data set.

One way of visualising grouped data is as a histogram. A histogram is a collection of rectangles, where the base of a rectangle (on the x-axis) covers the values in the range associated with it, and the height of a rectangle corresponds to the number of values in its range.

WORKED EXAMPLE 9: GROUPS AND HISTOGRAMS

QUESTION

The heights in centimetres of 30 learners are given below.

142	163	169	132	139	140	152	168	139	150
161	132	162	172	146	152	150	132	157	133
141	170	156	155	169	138	142	160	164	168

Group the data into the following ranges and draw a histogram of the grouped data:

$$130 \leq h < 140$$
$$140 \leq h < 150$$
$$150 \leq h < 160$$
$$160 \leq h < 170$$
$$170 \leq h < 180$$

(Note that the ranges do not overlap since each one starts where the previous one ended.)

WORKED EXAMPLE 9 CONTINUED: GROUPS AND HISTOGRAMS

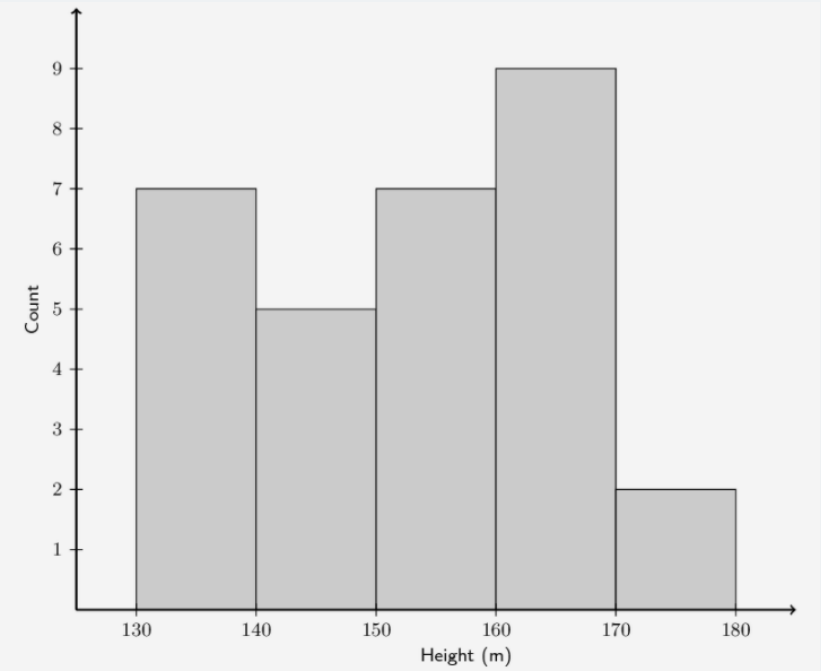
SOLUTION

Step 1: Count the number of values in each range

Range	Count
$130 \leq h < 140$	7
$140 \leq h < 150$	5
$150 \leq h < 160$	7
$160 \leq h < 170$	9
$170 \leq h < 180$	2

Step 2: Draw the histogram

Since there are 5 ranges, the histogram will have 5 rectangles. The base of each rectangle is defined by its range. The height of each rectangle is determined by the count in its range.



The histogram makes it easy to see in which range most of the heights are located and provides an overview of the distribution of the values in the data set.

3.1 Measures of central tendency

With grouped data our estimates of central tendency will change because we lose some information when we place each value in a range. If all we have to work with is the grouped data, we do not know the measured values to the same accuracy as before. The best we can do is to assume that values are grouped at the centre of each range.

Looking back to the previous worked example, we started with this data set of learners' heights.

$$\{132; 132; 132; 133; 138; 139; 139; 140; 141; 142; 142; 146; 150; 150; 152;$$
$$152; 155; 156; 157; 160; 161; 162; 163; 164; 168; 168; 169; 169; 170; 172\}$$

Note that the data are sorted.

The mean of these data is $151,8$ and the median is 152. The mode is 132, but remember that there are problems with computing the mode of continuous quantitative data.

After grouping the data, we now have the data set shown below. Note that each value is placed at the centre of its range and that the number of times that each value is repeated corresponds exactly to the counts in each range.

$$\{135; 135; 135; 135; 135; 135; 135; 145; 145; 145; 145; 145; 155; 155; 155;$$
$$155; 155; 155; 155; 165; 165; 165; 165; 165; 165; 165; 165; 175; 175\}$$

The grouping changes the measures of central tendency since each datum is treated as if it occurred at the centre of the range in which it was placed.

The mean is now 153, the median 155 and the mode is 165. This is actually a better estimate of the mode, since the grouping showed in which range the learners' heights were clustered.

NOTE

We can also just give the modal group and the median group for grouped data. The modal group is the group that has the highest number of data values. The median group is the central group when the groups are arranged in order.

4 MEASURES OF DISPERSION

The central tendency is not the only interesting or useful information about a data set. The two data sets illustrated below have the same mean (0), but have different spreads around the mean. Each circle represents one value from the data set (or one datum).

Dispersion is a general term for different statistics that describe how values are distributed around the centre. In this section we will look at measures of dispersion.

4.1 Range

DEFINITION

Range

The range of a data set is the difference between the maximum and minimum values in the set.

The most straightforward measure of dispersion is the range. The range simply tells us how far apart the largest and smallest values in a data set are. The range is very sensitive to outliers.

WORKED EXAMPLE 10: RANGE

QUESTION

Find the range of the following data set:

$$\{1; 4; 5; 8; 6; 7; 5; 6; 7; 4; 10; 9; 10\}$$

What would happen if we removed the first value from the set?

SOLUTION

Step 1: Determine the range

The smallest value in the data set is 1 and the largest value is 10.

The range is $10 - 1 = 9$

Step 2: Remove the first value

If the first value, 1, were to be removed from the set, the minimum value would be 4. This means that the range would change to $10 - 4 = 6$. 1 is not typical of the other values. It is an outlier and has a big influence on the range.

4.2 Percentiles

DEFINITION

Percentile

The p^{th} percentile is the value, v, that divides a data set into two parts, such that p percent of the values in the data set are less than v and $100 - p$ percent of the values are greater than v. Percentiles can lie in the range $0 \leq p \leq 100$.

To understand percentiles properly, we need to distinguish between 3 different aspects of a datum: its value, its rank and its percentile:

- The value of a datum is what we measured and recorded during an experiment or survey.
- The rank of a datum is its position in the sorted data set (for example, first, second, third, and so on).
- The percentile at which a particular datum is, tells us what percentage of the values in the full data set are less than this datum.

The table below summarises the value, rank and percentile of the data set:

$$\{14{,}2;\ 13{,}9;\ 19{,}8;\ 10{,}3;\ 13{,}0;\ 11{,}1\}$$

Value	Rank	Percentile
$10,3$	1	0
$11,1$	2	20
$13,0$	3	40
$13,9$	4	60
$14,2$	5	80
$19,8$	6	100

As an example, $13,0$ is at the 40^{th} percentile since there are 2 values less than $13,0$ and 3 values greater than $13,0$.

$$\frac{2}{2+3} = 0,4 = 40\%$$

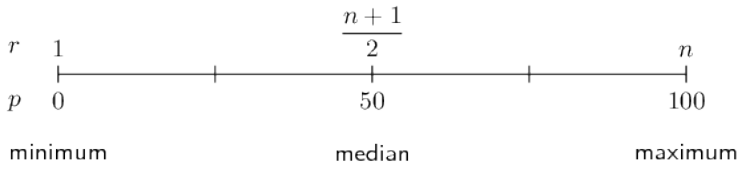
In general, the formula for finding the p^{th} percentile in an ordered data set with n values is

$$r = \frac{p}{100}(n-1) + 1$$

This gives us the rank, r, of the p^{th} percentile. To find the value of the p^{th} percentile, we have to count from the first value in the ordered data set up to the r^{th} value.

Sometimes the rank will not be an integer. This means that the percentile lies between two values in the data set. The convention is to take the value halfway between the two values indicated by the rank.

The figure below shows the relationship between rank and percentile graphically. We have already encountered three percentiles in this chapter: the median (50^{th} percentile), the minimum (0^{th} percentile) and the maximum (100^{th}). The median is defined as the value halfway in a sorted data set.



WORKED EXAMPLE 11: USING THE PERCENTILE FORMULA

QUESTION

Determine the minimum, maximum and median values of the following data set using the percentile formula.

$$\{14; 17; 45; 20; 19; 36; 7; 30; 8\}$$

SOLUTION

Step 1: Sort the values in the data set

Before we can use the rank to find values in the data set, we always have to order the values from the smallest to the largest. The sorted data set is

$$\{7; 8; 14; 17; 19; 20; 30; 36; 45\}$$

Step 2: Find the minimum

We already know that the minimum value is the first value in the ordered data set. We will now confirm that the percentile formula gives the same answer. The minimum is equivalent to the 0^{th} percentile. According to the percentile formula the rank, r, of the $p = 0^{\text{th}}$ percentile in a data set of $n = 9$ values is:

$$\begin{aligned} r &= \frac{p}{100}(n-1)+1 \\ &= \frac{0}{100}(9-1)+1 \\ &= 1 \end{aligned}$$

This confirms that the minimum value is the first value in the list, namely 7.

WORKED EXAMPLE 11 CONTINUED: USING THE PERCENTILE FORMULA

Step 3: Find the maximum

We already know that the maximum value is the last value in the ordered data set. The maximum is also equivalent to the 100^{th} percentile. Using the percentile formula with $p = 100$ and $n = 9$, we find the rank of the maximum value is:

$$\begin{aligned} r &= \frac{p}{100}(n-1)+1 \\ &= \frac{100}{100}(9-1)+1 \\ &= 9 \end{aligned}$$

This confirms that the maximum value is the last (the ninth) value in the list, namely 45.

Step 4: Find the median

The median is equivalent to the 50^{th} percentile. Using the percentile formula with $p = 50$ and $n = 9$, we find the rank of the median value is:

$$\begin{aligned} r &= \frac{50}{100}(n-1)+1 \\ &= \frac{50}{100}(9-1)+1 \\ &= \frac{1}{2}(8)+1 \\ &= 5 \end{aligned}$$

This shows that the median is in the middle (at the fifth position) of the ordered data set. Therefore the median value is 19.

DEFINITION

Quartiles

The quartiles are the three data values that divide an ordered data set into four groups, where each group contains an equal number of data values. The median (50^{th} percentile) is the second quartile ($Q2$). The 25^{th} percentile is also called the first or lower quartile ($Q1$). The 75^{th} percentile is also called the third or upper quartile ($Q3$).

WORKED EXAMPLE 12: QUARTILES

QUESTION

Determine the quartiles of the following data set:

$$\{7; 45; 11; 3; 9; 35; 31; 7; 16; 40; 12; 6\}$$

SOLUTION

Step 1: Sort the data set

$$\{3; 6; 7; 7; 9; 11; 12; 16; 31; 35; 40; 45\}$$

Step 2: Find the ranks of the quartiles

Using the percentile formula with $n = 12$, we can find the rank of the 25^{th}, 50^{th} and 75^{th} percentiles:

$$\begin{aligned} r_{25} &= \frac{25}{100}(12-1)+1 \\ &= 3,75 \\ r_{50} &= \frac{50}{100}(12-1)+1 \\ &= 6,5 \\ r_{75} &= \frac{75}{100}(12-1)+1 \\ &= 9,25 \end{aligned}$$

Step 3: Find the values of the quartiles

Note that each of these ranks is a fraction, meaning that the value for each percentile is somewhere in between two values from the data set.

For the 25^{th} percentile the rank is $3,75$, which is between the third and fourth values. Since both these values are equal to 7, the 25^{th} percentile is 7.

For the 50^{th} percentile (the median) the rank is $6,5$, meaning halfway between the sixth and seventh values. The sixth value is 11 and the seventh value is 12, which means that the median is $\frac{11+12}{2} = 11,5$. For the 75^{th} percentile the rank is $9,25$, meaning between the ninth and tenth values. Therefore the 75^{th} percentile is $\frac{31+35}{2} = 33$.

Deciles

The deciles are the nine data values that divide an ordered data set into ten groups, where each group contains an equal number of data values. For example, consider the ordered data set:

$$28; 33; 35; 45; 57; 59; 61; 68; 69; 72; 75; 78; 80; 83; 86; 91;$$
$$92; 95; 101; 105; 111; 117; 118; 125; 127; 131; 137; 139; 141$$

The nine deciles are: $35; 59; 69; 78; 86; 95; 111; 125; 137$.

4.3 Percentiles for grouped data

In grouped data, the percentiles will lie somewhere inside a range, rather than at a specific value. To find the range in which a percentile lies, we still use the percentile formula to determine the rank of the percentile and then find the range within which that rank is.

WORKED EXAMPLE 13: PERCENTILES IN GROUPED DATA

QUESTION

The mathematics marks of 100 grade 10 learners at a school have been collected. The data are presented in the following table:

Percentage mark	Number of learners
$0 \leq x < 20$	2
$20 \leq x < 30$	5
$30 \leq x < 40$	18
$40 \leq x < 50$	22
$50 \leq x < 60$	18
$60 \leq x < 70$	13
$70 \leq x < 80$	12
$80 \leq x < 100$	10

1. Calculate the mean of this grouped data set.
2. In which intervals are the quartiles of the data set?
3. In which interval is the 30^{th} percentile of the data set?

SOLUTION

Step 1: Calculate the mean

Since we are given grouped data rather than the original ungrouped data, the best we can do is approximate the mean as if all the learners in each interval were located at the central value of the interval.

$$\text{Mean } = \frac{2(10) + 5(25) + 18(35) + 22(45) + 18(55) + 13(65) + 12(75) + 10(90)}{100} = 54\%$$

WORKED EXAMPLE 13 CONTINUED: PERCENTILES IN GROUPED DATA

Step 2: Find the quartiles

Since the data have been grouped, they have also already been sorted. Using the percentile formula and the fact that there are 100 learners, we can find the rank of the 25^{th}, 50^{th} and 75^{th} percentiles as

$$\begin{aligned} r_{25} &= \frac{25}{100}(100-1)+1 \\ &= 24,75 \\ r_{50} &= \frac{50}{100}(100-1)+1 \\ &= 50,5 \\ r_{75} &= \frac{75}{100}(100-1)+1 \\ &= 75,25 \end{aligned}$$

Now we need to find in which ranges each of these ranks lie.

- For the lower quartile, we have that there are $2+5=7$ learners in the first two ranges combined and $2+5+18=25$ learners in the first three ranges combined. Since $7 < r_{25} < 25$, this means the lower quartile lies somewhere in the third range: $30 \leq x < 40$.
- For the second quartile (the median), we have that there are $2+5+18+22=47$ learners in the first four ranges combined. Since $47 < r_{50} < 65$, this means that the median lies somewhere in the fifth range: $50 \leq x < 60$.
- or the upper quartile, we have that there are 65 learners in the first five ranges combined and $65+13=78$ learners in the first six ranges combined. Since $65 < r_{75} < 78$, this means that the upper quartile lies somewhere in the sixth range: $60 \leq x < 70$.

Step 3: Find the 30^{th} percentile

Using the same method as for the quartiles, we first find the rank of the 30^{th} percentile.

$$\begin{aligned} r &= \frac{30}{100}(100-1)+1 \\ &= 30,7 \end{aligned}$$

Now we have to find the range in which this rank lies. Since there are 25 learners in the first 3 ranges combined and 47 learners in the first 4 ranges combined, the 30^{th} percentile lies in the fourth range: $40 \leq x < 50$

4.4 Ranges

We define data ranges in terms of percentiles. We have already encountered the full data range, which is simply the difference between the 100^{th} and the 0^{th} percentile (that is, between the maximum and minimum values in the data set).

DEFINITION

Interquartile range

The interquartile range is a measure of dispersion, which is calculated by subtracting the first quartile ($Q1$) from the third quartile ($Q3$). This gives the range of the middle half of the data set.

DEFINITION

Semi interquartile range

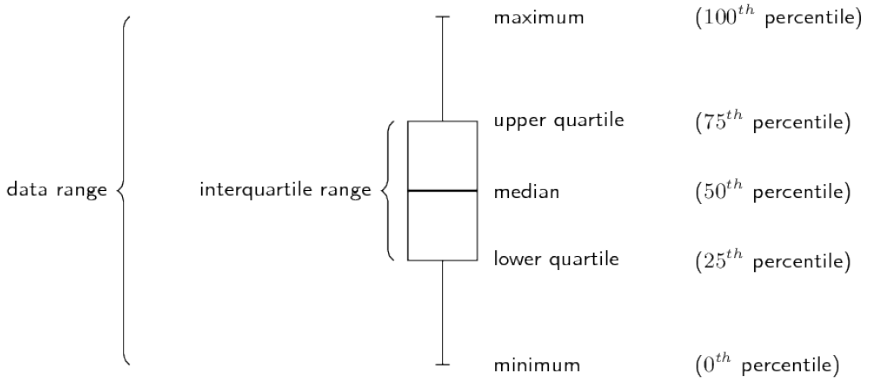
The semi interquartile range is half of the interquartile range.

5 FIVE NUMBER SUMMARY

A common way of summarising the overall data set is with the five number summary and the box-and-whisker plot. These two represent exactly the same information, numerically in the case of the five number summary and graphically in the case of the box-and-whisker plot.

The five number summary consists of the minimum value, the maximum value and the three quartiles. Another way of saying this is that the five number summary consists of the following percentiles: 0^{th}, 25^{th}, 50^{th}, 75^{th}, 100^{th}.

The box-and-whisker plot shows these five percentiles as in the figure below. The box shows the interquartile range (the distance between $Q1$ and $Q3$). A line inside the box shows the median. The lines extending outside the box (the whiskers) show where the minimum and maximum values lie. This graph can also be drawn horizontally.



WORKED EXAMPLE 14: FIVE NUMBER SUMMARY

QUESTION

Draw a box and whisker diagram for the following data set:

$$\{1{,}25;\ 1{,}5;\ 2{,}5;\ 2{,}5;\ 3{,}1;\ 3{,}2;\ 4{,}1;\ 4{,}25;\ 4{,}75;\ 4{,}8;\ 4{,}95;\ 5{,}1\}$$

SOLUTION

Step 1: Determine the minimum and maximum

Since the data set is already sorted, we can read off the minimum as the first value $(1, 25)$ and the maximum as the last value $(5, 1)$.

Step 2: Determine the quartiles

There are 12 values in the data set. Using the percentile formula, we can determine that the median lies between the sixth and seventh values, making it:

$$\frac{3,2 + 4,1}{2} = 3,65$$

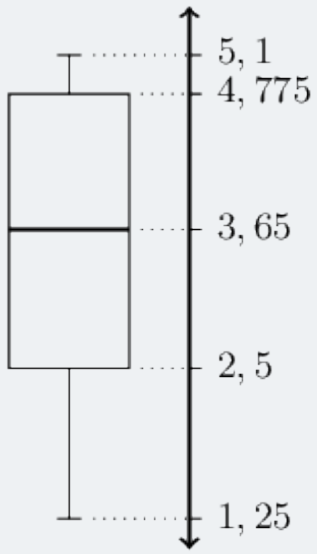
The first quartile lies between the third and fourth values, making it:

$$\frac{2,5 + 2,5}{2} = 2,5$$

The third quartile lies between the ninth and tenth values, making it:

$$\frac{4,75 + 4,8}{2} = 4,775$$

This provides the five number summary of the data set and allows us to draw the following box-and-whisker plot.



6 CHAPTER SUMMARY

- Data refer to the pieces of information that have been observed and recorded, from an experiment or a survey.
- Quantitative data are data that can be written as numbers. Quantitative data can be discrete or continuous.
- Qualitative data are data that cannot be written as numbers. There are two common types of qualitative data: categorical and anecdotal data.
- The mean is the sum of a set of values divided by the number of values in the set.

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{x_1 + x_2 + \cdots + x_n}{n}\end{aligned}$$

- The median of a data set is the value in the central position, when the data set has been arranged from the lowest to the highest value. If there are an odd number of data, the median will be equal to one of the values in the data set. If there are an even number of data, the median will lie half way between two values in the data set.
- The mode of a data set is the value that occurs most often in the set.
- An outlier is a value in the data set that is not typical of the rest of the set. It is usually a value that is much greater or much less than all the other values in the data set.
- Continuous quantitative data can be grouped by dividing the full range of values into a few sub-ranges. By assigning each continuous value to the sub-range or class within which it falls, the data set changes from continuous to discrete.
- Dispersion is a general term for different statistics that describe how values are distributed around the centre.
- The range of a data set is the difference between the maximum and minimum values in the set.
- The p^{th} percentile is the value, v, that divides a data set into two parts, such that p% of the values in the data set are less than v and $100 - p$% of the values are greater than v. The general formula for finding the p^{th} percentile in an ordered data set of n values is

$$r = \frac{p}{100}(n-1) + 1$$

- The quartiles are the three data values that divide an ordered data set into four groups, where each group contains an equal number of data values. The lower quartile is denoted $Q1$, the median is $Q2$ and the upper quartile is $Q3$.

- The interquartile range is a measure of dispersion, which is calculated by subtracting the lower (first) quartile from the upper (third) quartile. This gives the range of the middle half of the data set.
- The semi interquartile range is half of the interquartile range.
- The five number summary consists of the minimum value, the maximum value and the three quartiles ($Q1$, $Q2$ and $Q3$).
- The box-and-whisker plot is a graphical representation of the five number summary.

7 EXERCISES

7.1 Exercise 1

1. The following data set of dreams that learners have was collected from Grade 12 learners just after their final exams.

   "I want to build a bridge!" ; "I want to help the sick." ; "I want running water!"

   Categorize the data set.

2. Calculate the mean of the following data set:

   $$\{9; 14; 9; 14; 8; 8; 9; 8; 9; 9\}$$

   Round your answer to 1 decimal place.

3. The following data set of sweets in a packet was collected from visitors to a sweet shop.

   $$\{23; 25; 22; 26; 27; 25; 21; 28\}$$

   Categorize the data set.

4. Calculate the median of the following data set:

   $$\{4; 13; 10; 13; 13; 4; 2; 13; 13; 13\}$$

5. The following data set of questions answered correctly was collected from a class of maths learners.

   $$\{3; 5; 2; 6; 7; 5; 1; 2\}$$

   Categorize the data set.

6. Calculate the mode of the following data set:

   $$\{6; 10; 6; 6; 13; 12; 12; 7; 13; 6\}$$

7. Calculate the mean, median and mode of the following data sets:

   7.1 $\{2; 5; 8; 8; 11; 13; 22; 23; 27\}$

   7.2 $\{15; 17; 24; 24; 26; 28; 31; 43\}$

   7.3 $\{4; 11; 3; 15; 11; 13; 25; 17; 2; 11\}$

   7.4 $\{24; 35; 28; 41; 31; 49; 31\}$

8. The ages of 15 runners of the Comrades Marathon were recorded:

   $$\{31; 42; 28; 38; 45; 51; 33; 29; 42; 26; 34; 56; 33; 46; 41\}$$

   Calculate the mean, median and modal age.

9. A group of 10 friends each have some stones. They work out that the mean number of stones they have is 6. Then 7 friends leave with an unknown number (x) of stones. The remaining 3 friends work out that the mean number of stones they have left is $12,33$.

   When the 7 friends left, how many stones did they take with them?

10. A group of 9 friends each have some coins. They work out that the mean number of coins they have is 4. Then 5 friends leave with an unknown number (x) of coins. The remaining 4 friends work out that the mean number of coins they have left is $2,5$.

    When the 5 friends left, how many coins did they take with them?

11. A group of 9 friends each have some marbles. They work out that the mean number of marbles they have is 3. Then 3 friends leave with an unknown number (x) of marbles. The remaining 6 friends work out that the mean number of marbles they have left is $1,17$.

    When the 3 friends left, how many marbles did they take with them?

12. In the first of a series of jars, there is 1 sweet. In the second jar, there are 3 sweets. The mean number of sweets in the first two jars is 2.

    12.1 If the mean number of sweets in the first three jars is 3 , how many sweets are there in the third jar?

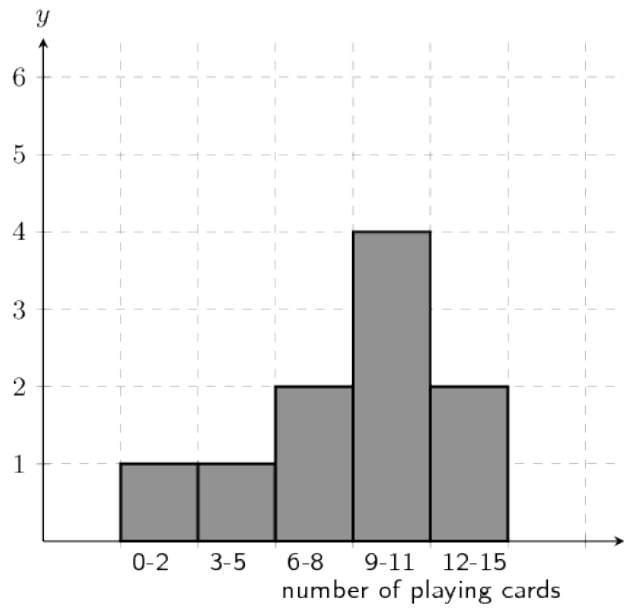
    12.2 If the mean number of sweets in the first four jars is 4 , how many sweets are there in the fourth jar?

13. Find a set of five ages for which the mean age is 5, the modal age is 2 and the median age is 3 years.

14. Four friends each have some marbles. They work out that the mean number of marbles they have is 10. One friend leaves with 4 marbles. How many marbles do the remaining friends have together?

7.2 Exercise 2

1. A group of 10 learners count the number of playing cards they each have. This is a histogram describing the data they collected:

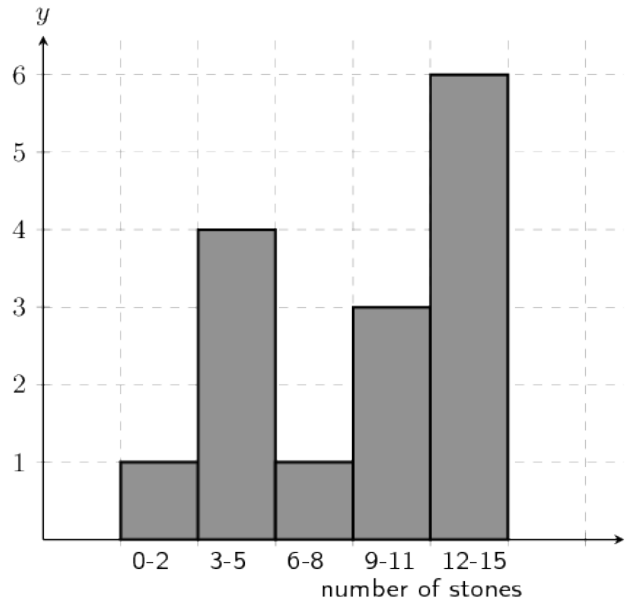


   Count the number of playing cards in the following range: $0 \leq$ number of playing cards ≤ 2

2. Consider the following grouped data and calculate the mean, the modal group and the median group.

Mass (kg)	Count
$40 < m < 45$	7
$45 < m \leq 50$	10
$50 < m \leq 55$	15
$55 < m \leq 60$	12
$60 < m \leq 65$	6

3. A group of 15 learners count the number of stones they each have. This is a histogram describing the data they collected:



Count the number of stones in the following range: $0 \leq$ number of stones ≤ 2

4. Find the mean, the modal group and the median group in this data set of how much time people needed to complete a game.

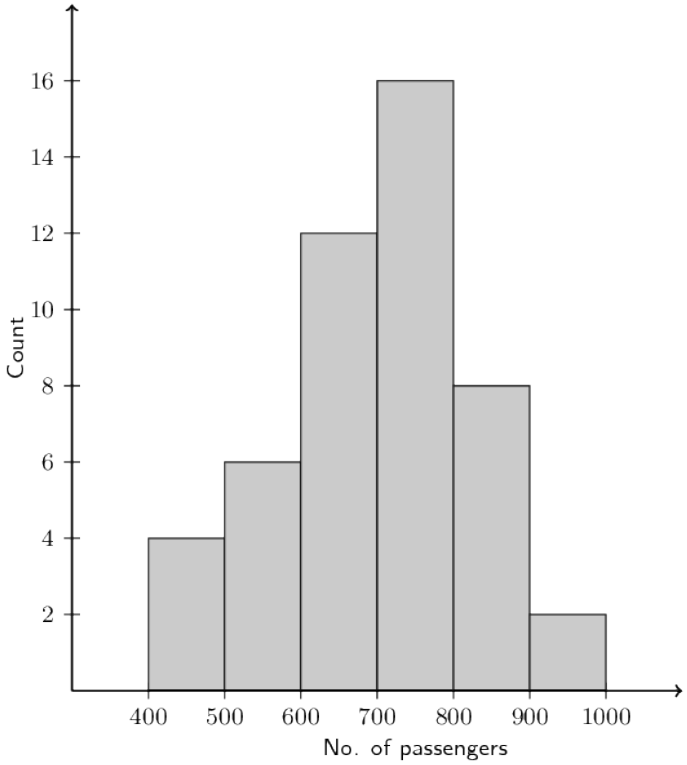
Time(s)	Count
$35 < t \leq 45$	5
$45 < t \leq 55$	11
$55 < t \leq 65$	15
$65 < t \leq 75$	26
$75 < t \leq 85$	19
$85 < t \leq 95$	13
$95 < t \leq 105$	6

5. A group of 20 learners count the number of playing cards they each have.
   This is the data they collect:

14	9	11	8	13
2	3	4	16	17
9	19	10	14	4
6	16	11	2	17

   Count the number of learners who have from 12 up to 15 playing cards. In other words, how many learners have playing cards in the following range: $12 \leq$ number of playing cards ≤ 15?
   It may be helpful for you to draw a histogram in order to answer the question.

6. The histogram below shows the number of passengers that travel in Alfred's minibus taxi per week.



Calculate:

6.1 The modal interval

6.2 The total number of passengers to travel in Alfred's taxi

6.3 An estimate of the mean

6.4 An estimate of the median

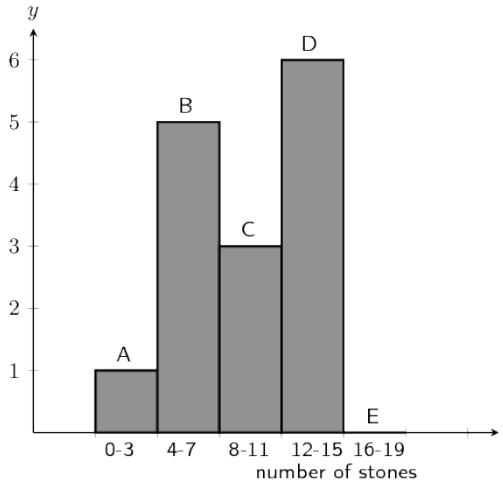
6.5 If it is estimated that every passenger travelled an average distance of 5km, how much money would Alfred have made if he charged R$3,50$ per km?

7. A group of 20 learners count the number of stones they each have. This is the data they collect:

16	6	11	19	20
17	13	1	5	12
5	2	16	11	16
6	10	13	6	17

Count the number of learners who have from 4 up to 7 stones. In other words, how many learners have stones in the following range: $4 \leq$ number of stones ≤ 7?
It may be helpful for you to draw a histogram in order to answer the question.

8. A group of 20 learners count the number of stones they each have. The learners draw a histogram describing the data they collected. However, they have made a mistake in drawing the histogram.

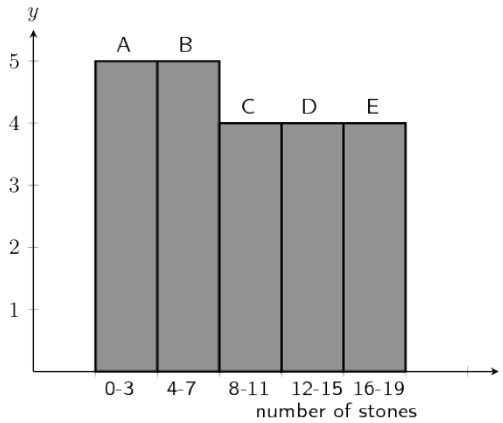


The data set below shows the correct information for the number of stones the learners have. Each value represents the number of stones for one learner.

$$\{4; 12; 15; 14; 18; 12; 17; 15; 1; 6; 6; 12; 6; 8; 6; 8; 17; 19; 16; 8\}$$

Help them figure out which column in the histogram is incorrect.

9. A group of 20 learners count the number of stones they each have. The learners draw a histogram describing the data they collected. However, they have made a mistake in drawing the histogram.

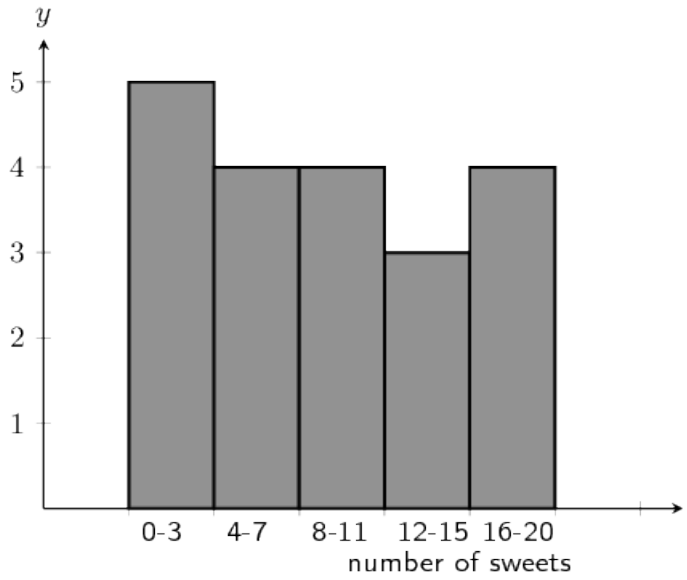


The data set below shows the correct information for the number of stones the learners have.

$$\{19; 11; 5; 2; 3; 4; 14; 2; 12; 19; 11; 14; 2; 19; 11; 5; 17; 10; 1; 12\}$$

Help them figure out which column in the histogram is incorrect.
Each value represents the number of stones for one learner.

10. A group of learners count the number of sweets they each have. This is a histogram describing the data they collected:



A cat jumps onto the table, and all their notes land on the floor, mixed up, by accident!
Help them find which of the following data sets match the above histogram:

Data set A:

2	1	20	10	5
3	10	2	6	1
2	2	17	3	18
3	7	10	8	18

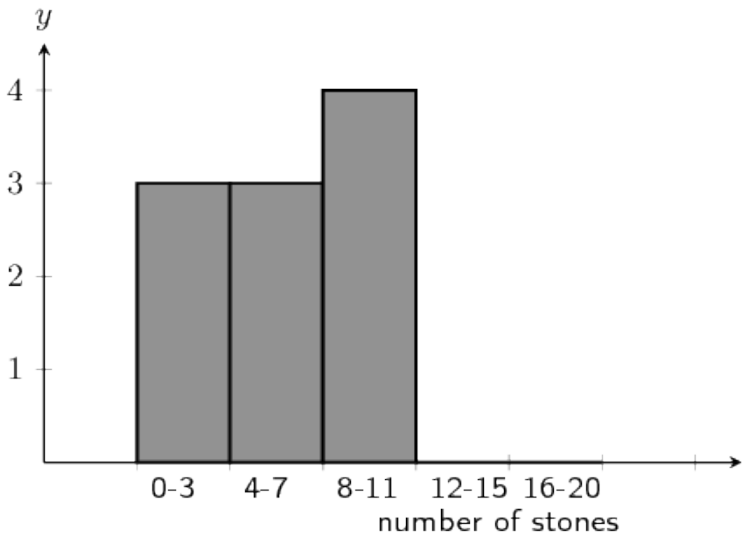
Data set B:

2	9	12	10	5
9	9	10	13	6
5	11	10	7	7

Data set C:

3	12	16	10	15
17	18	2	3	7
11	12	8	2	7
17	3	11	4	4

11. A group of learners count the number of stones they each have. This is a histogram describing the data they collected:



A cleaner knocks over their table, and all their notes land on the floor, mixed up, by accident! Help them find which of the following data sets match the above histogram:

Data set A:

12	4	2	15	10
18	10	16	16	19
1	2	9	10	16
10	11	9	2	13

Data set B:

7	10	4	5	8
7	12	10	14	5
1	9	2	13	3

Data set C:

9	3	8	5	8
5	8	1	4	3

7.3 Exercise 3

1. Lisa is working in a computer store. She sells the following number of computers each month:

$$\{27; 39; 3; 15; 43; 27; 19; 54; 65; 23; 45; 16\}$$

   Give the five number summary and box-and-whisker plot of Lisa's sales.

2. Zithulele works as a telesales person. He keeps a record of the number of sales he makes each month. The data below show how much he sells each month.

$$\{49; 12; 22; 35; 2; 45; 60; 48; 19; 1; 43; 12\}$$

   Give the five number summary and box-and-whisker plot of Zithulele's sales.

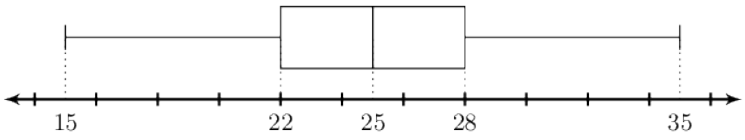
3. Nombusa has worked as a florist for nine months. She sold the following number of wedding bouquets:

$$\{16; 14; 8; 12; 6; 5; 3; 5; 7\}$$

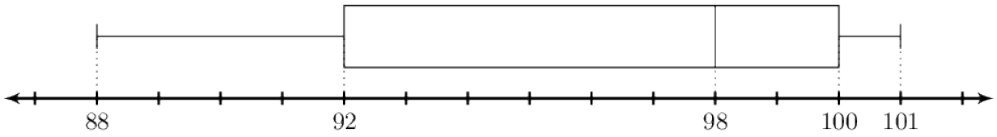
   Give the five number summary of Nombusa's sales.

4. Determine the five number summary for each of the box-and-whisker plots below.

   4.1 The plot:



   4.2 The plot:



7.4 Exercise 4

1. A group of 15 learners count the number of sweets they each have. This is the data they collect:

4	11	14	7	14
5	8	7	12	12
5	13	10	6	7

   Calculate the range of values in the data set.

2. A group of 10 learners count the number of playing cards they each have. This is the data they collect:

5	1	3	1	4
10	1	3	3	4

   Calculate the range of values in the data set.

3. Find the range of the data set

$$\{1; 2; 3; 4; 4; 4; 5; 6; 7; 8; 8; 9; 10; 10\}$$

4. What are the quartiles of this data set?

$$\{3; 5; 1; 8; 9; 12; 25; 28; 24; 30; 41; 50\}$$

5. A class of 12 learners writes a test and the results are as follows:

$$\{20; 39; 40; 43; 43; 46; 53; 58; 63; 70; 75; 91\}$$

   Find the range, quartiles and the interquartile range.

6. Three sets of data are given:

   Data set 1:

$$\{9; 12; 12; 14; 16; 22; 24\}$$

   Data set 2:

$$\{7; 7; 8; 11; 13; 15; 16\}$$

   Data set 3:

$$\{11; 15; 16; 17; 19; 22; 24\}$$

   For each data set find:

   6.1 The range

   6.2 The lower quartile

   6.3 The median

   6.4 The upper quartile

   6.5 The interquartile range

   6.6 The semi-interquartile range

8 ANSWERS TO EXERCISES

8.1 Exercise 1

1. The data set is qualitative anecdotal.
2. Mean : $9,7$
3. The data set is quantitative discrete.
4. Median : 13
5. The data set is quantitative discrete.
6. Mode : 6

7.1 Mean : $13,2$
Median : 11
Mode : 8

7.2 Mean : 26
Median : 25
Mode : 24

7.3 Mean : $11,2$
Median : 11
Mode : 11

7.4 Mean : $34,3$
Median : 31
Mode : 31

8. Mean : $38,3$
Median : 31
Mode : 32 and 42
9. They took 23 stones.
10. They took 26 coins.
11. They took 2 marbles.

12.1 5 sweets

12.2 7 sweets

13. Data set 1: $\{2; 2; 3; 4; 14\}$
    Data set 2: $\{2; 2; 3; 5; 13\}$
    Data set 3: $\{2; 2; 3; 6; 12\}$
    Data set 4: $\{2; 2; 3; 7; 11\}$
    Data set 5: $\{2; 2; 3; 8; 10\}$

14. 36 marbles

8.2 Exercise 2

1. 1 playing card

2. Mean : 52
   Modal group : $50 < m \leq 55$
   Median group : $50 < m \leq 55$

3. 1 stone

4. Mean : $70,66$
   Modal group : $65 < t \leq 75$
   Median group : $465 < t \leq 75$

5. 3 learners

6.1 $700 < x \leq 800$ with 16 values.

6.2 $33\,600$ passengers

6.3 Mean : 700

6.4 Median : 750

6.5 R$588\,000$

7. 5 learners

8. The column with the error in it was: E.

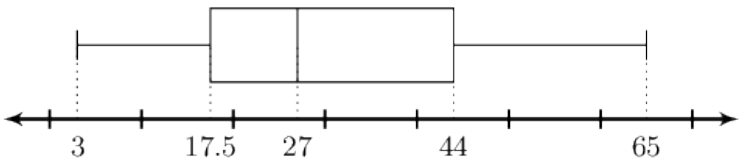
9. The column with the error in it was: B.

10. The correct answer is: Data Set C

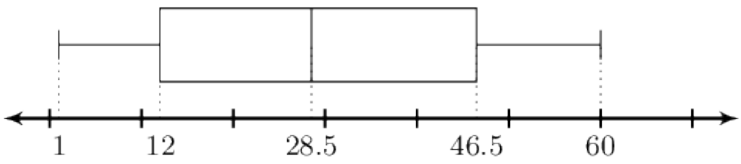
11. The correct answer is: Data Set C

8.3 Exercise 3

1. Minimum : 3
   $Q_1; 17,5$
   Median : 27
   Q_3 : 44
   Maximum : 65



2. Minimum: 1
   Q_1 : 12
   Median : $28,5$
   Q_3 : $46,5$
   Maximum: 60



3. Minimum: 3
   Q_1 : 5
   Median : 7
   Q_3 : 12
   Maximum: 16

4.1 Minimum: 15
   Q_1 : 22
   Median: 25
   Q_3 : 28
   Maximum: 35

4.2 Minimum: 88
   Q_1 : 92
   Median: 98
   Q_3 : 100
   Maximum: 101

8.4 Exercise 4

1. Range : 10
2. Range : 9
3. Range : 9
4. 25^{th} percentile : $6,5$
   Median : 18
   75^{th} percentile : 29
5. Range $= 71$
   $25^{th} = 41,5$
   $50^{th} = 49,5$
   $75^{\text{th}} = 66,5$
   Interquartile range $= 25$

6.1 Data set 1 : 15
Data set 2 : 9
Data set 3 : 13

6.2 Data set 1 : 12
Data set 2 : $7,5$
Data set 3 : $15,5$

6.3 Data set 1 : 14
Data set 2 : 11
Data set 3 : 17

6.4 Data set 1 : 19
Data set 2 : 14
Data set 3 : 20.5

6.5 Data set 1 : 7
Data set 2 : $6,5$
Data set 3 : 5

6.6 Data set 1 : $3,5$
Data set 2 : $3,25$
Data set 3 : $2,5$