SECTION 9

STATISTICS

HANDLING DATA

Organising, representing and interpreting data

INTRODUCTION

Do you keep monitoring how many likes your social media posts have at certain points in the day or week? Or how many followers you have received after a particular post? The data you gathered from monitoring the interaction with your posts helps you decide on what content to keep putting out on social media and that is what we call decision-making. You did not make this decision out of the blue but from the data you received based on the interactions with your posts. That is just a pinch of how statistics influences our everyday lives. Having statistical knowledge helps in tracking rainfall patterns to make good decisions in farming and monitoring our nutrient intake so we practise healthy eating habits among others. In this section, we examine how data is collected, organised and categorised. We will use graphs and charts to represent these continuous changes, revealing trends and patterns over time. Measures of central tendency and dispersion will be elaborated on. Throughout this section, we will explore real-life applications relevant to Ghana. From analysing market trends to understanding crop yields, you will gain invaluable skills to make informed choices for yourself and your community.

At the end of this section, you will be able to:

- Identify and present appropriate ways of collecting and representing data.
- Categorise data and determine which scale of measurement describes the data.
- Organise data into appropriate frequency distribution tables manually and with Microsoft Excel.
- Present data using appropriate graphs by hand and/or by technology and justify why a particular representation is more suitable than others for a given situation.
- Present grouped data using appropriate graphs by hand and/or by appropriate technology and justify why a particular representation is more suitable than others for a given situation.

- Compare various statistical representations and justify why a particular representation is more suitable than others for a given situation.
- Calculate measures of central tendencies (mode, mean and median) for a given data by formulas or other techniques and establish which is appropriate to report on a given data.
- Work out simple measures of dispersion (range, quartile, inter-quartile etc) for raw data and interpret them in context.

Key Ideas:

- Data can be collected through interviews, questionnaires and observations.
- Collected data can be represented through tables and graphs.
- Sampling helps us choose a representative group to learn about a larger population.
- Data can be quantitative or qualitative.
- Categorical data can be represented using bar charts and pie charts.
- Continuous data can be represented using line graphs, histograms and cumulative frequency curves (ogives).
- The measure of central tendencies includes mean, mode and median.
- The measures of dispersion include range, standard deviation, variance and interquartile range.

COLLECTING AND REPRESENTING DATA

Have you ever heard the statement "*Tell me about yourself*"? What usually follows is a detailed description of who the person is, their interests, goals and more. Whoever asked "Tell me about yourself" now leaves with data on the person who responded to the statement and, as such, has *collected* data. Data can be collected in different ways, let's explore some of the ways through these activities.

Activity 9.1

1. Pick a classmate.
2. Ask them who they think should lead form/class meetings and record their response.
3. Ask them to share some of the things they think should be discussed in form/class meetings. Record their response.
4. Repeat steps 1 – 3 for about five people.
5. Take some time to go over the recorded responses.
6. Decide what can be done based on those responses.
7. What is the means by which you collected this data?

It is an **interview** because you had one-on-one discussions asking targeted questions.

Activity 9.2

1. Choose a topic you want to investigate (e.g., favourite hobbies, study habits or opinions on a school policy).
2. Design a 5-7 question survey. This is called a **questionnaire**.

   E.g. Study Habits.

Table 1: A Sample questionnaire

Item	Always	Often	Sometimes	Never
I review my notes before an examination				
I use practice tests to prepare for exams				
I set specific study goals before studying				
I feel that I have enough time to prepare for exams or complete assignments				
I study with distractions (e.g. TV, music, phone, etc.) nearby.				

3. Distribute the questionnaire to at least 10 peers.
4. Collect the responses and record them systematically.
5. Summarise the data, identify trends, and create simple graphs (bar charts, pie charts) to present your findings.
6. Share the results with a friend and discuss.

Activity 9.3

1. Choose a peer performing a particular activity (ironing, washing, leading a presentation, carrying out a science experiment in the laboratory, performing at a variety show during entertainment, etc.)
2. Pay attention to everything this peer does in carrying out the activity.
3. As you watch the activity being carried out, write down what you see.

Prompts for documenting

1. What was done first?
2. Followed by what?
3. And then?

This data has been collected through **observation**.

Sampling

During data collection, you may not be able to interview, observe or provide a questionnaire for everyone hence choosing a representative group to learn about the larger population of interest is recommended. This process of choosing the representative group is what is termed "sampling". There are many techniques for sampling depending on what you intend to do with the data. Some common sampling techniques include:

a. **Simple Random Sampling:** this is a sampling method where each member of a population has an equal chance of being selected. It is a type of probability sampling where participants are chosen entirely by chance, ensuring that every possible sample combination has the same likelihood of being selected.

Activity 9.4

1. Write the names of all students in your form on different pieces of paper.
2. Put all the pieces of paper in a bag and shuffle them well.
3. Ask one student to randomly draw 6 pieces of paper one after the other and call out the names on them.
4. The 6 people named represent the randomly selected sample.

b. Systematic Sampling: this is a probability sampling method where you select samples from a population using a fixed, periodic interval. You choose every k^{th} member from a list after selecting a random starting point. This method is often used when a population is ordered or when you want a straightforward sampling process.

Activity 9.5

1. Go to the novel section of your school library and create a list of the first 35 books you come across.
2. Now, the goal is to select 7 books from the list of 35.
3. To calculate the interval (k), divide 35 by 7 and document the answer.
4. From your list, pick a random starting point between 1 and k (e.g. 4)
5. From the random starting point (e.g. 4), select every k^{th} (5th) item until you have 7 books.
6. Record the titles of the books and, with a classmate, discuss your results.

c. Stratified Sampling: this is a probability sampling technique where a population is divided into smaller subgroups (strata) based on a key characteristic (e.g., gender, year group, class, region, age group, etc.) After dividing the population, randomly select participants from each subgroup to ensure your sample reflects the population proportions.

Activity 9.6

Create a sample of 20 learners who prefer a particular subject from a list of 40 students in total.

1. Create a list of learners who prefer a particular subject in the options

(E.g, English - 4, Mathematics - 6, Science - 18, Social Studies -12).

2. One person can be in only one group hence we have 4 subgroups.
3. Determine how many learners to sample from each subgroup based on their proportion in the total population. So in our example:

   English: $\frac{4}{40} \times 20 = 2$

   Mathematics: $\frac{6}{40} \times 20 = 3$

   Science: $\frac{18}{40} \times 20 = 9$

   Social studies: $\frac{12}{40} \times 20 = 6$
4. Now, randomly select 2 people from those who prefer English, 3 from those who prefer mathematics, 9 from those who prefer science and 6 from those who prefer social studies.

You have now created a sample of 20 learners with each preferred subject adequately represented.

d. **Cluster Sampling** is a probability sampling method where the population is divided into distinct groups or "clusters" and a random selection of clusters is carried out. Instead of sampling individuals from the entire population, all or a random selection of individuals from the chosen clusters are used. It is useful in geographically spread-out populations or when reaching individuals within groups is easier (e.g., studying health practices in randomly chosen villages).

e. **Convenience Sampling** is a non-probability sampling method where participants are selected based on their availability, accessibility and ease of participation. You choose the easiest people to access (e.g., classmates, neighbours). This method is not ideal for generalisable research as the results may not be representative of the whole population.

f. **Purposive Sampling** (Judgmental Sampling) is a non-probability sampling method where participants are selected based on specific characteristics needed by the researcher to carry out research. It is ideal for in-depth information from a specific group with expertise or experience (e.g., interviewing experienced cocoa farmers about sustainable practices).

g. **Snowball Sampling** is a non-probability sampling method where existing participants recruit future participants from their acquaintances or social networks. It is often used in studies where the population is hard to reach or

identify, such as hidden, sensitive or specialised groups (e.g., people with rare diseases, drug users or specific professionals).

The sampling process starts with a few participants who meet the criteria and then refer others in their network who share similar characteristics. It is useful for studying hard-to-reach populations.

h. **Quota Sampling** is a non-probability sampling method where the researcher divides the population into specific subgroups (or "quotas") and selects a certain number of participants from each subgroup to ensure the sample reflects the characteristics of the population. Unlike probability methods, participants within each subgroup are selected based on convenience or other non-random criteria and not proportionately.

The best sampling technique depends on the research question and the population you are studying.

Generalisation:

- Data can be collected through interviews, questionnaires and observations.
- Data can be represented using tables and graphs.

DATA CATEGORISATION

The data we gather about people, places, objects, etc. come in different forms. It could be numerical like age, height, weight or categorical like colours, complexion, brands of cars, model of phones, Region of origin, telecommunication network used, etc. Considering this, data is categorised into two main types: qualitative and quantitative.

Quantitative Data: This involves data that is measurable and expressed in numerical form. It represents quantities or amounts and can be subjected to mathematical operations like addition and subtraction.

Qualitative Data: This involves non-numerical information that describes the characteristics, qualities, or attributes of something. It is often descriptive and can be observed but not measured in numbers. Qualitative data typically involves text, images or categories and is used to capture complex details about experiences, behaviours or opinions. It is more about words and stories.

Activity 9.7

1. Create a table with two columns labelled qualitative and quantitative.
2. From the list of characteristics of interest provided, decide which ones are qualitative or quantitative and group them in the table accordingly.

   List of characteristics:

   a. Number of oranges in a basket
   b. Country of origin
   c. Height of students in Form 1
   d. Weight of students in Home Economics class
   e. Brands of rice
   f. Designers of bags
   g. Gender of a student
   h. Football teams in Ghana
   i. Weather temperature
3. Now, list your characteristics of interest and group them under quantitative or qualitative.
4. Compare and discuss your results with a classmate.

Categorisation of qualitative and quantitative data

Quantitative data may be continuous or discrete.

1. **Discrete data** refers to numerical data that can only take specific, distinct values. These values are countable and typically represent whole numbers. Discrete data often arises from counting processes with no intermediate values. Such data include the number of light bulbs in a school, the number of desks, the number of learners in an auditorium, etc.
2. **Continuous data** refers to numerical data that can take any value (real number) within a given range. Unlike discrete data, continuous data can represent measurements that are not limited to specific, separate values but can fall anywhere on a continuous scale (on the number line). Examples include age, weight, height or temperature of learners, etc.

Levels of Measurement

Qualitative and quantitative data are further placed into four levels or scales of measurement namely; nominal, ordinal, interval and ratio. Measurement scales tell us what kind of information the numbers we assign represent.

1. ***Nominal*** – this is a qualitative, categorical measurement scale used to classify data into distinct categories or groups without implying any order or ranking among them. It is the simplest level of measurement and is used to name or label variables where the categories are mutually exclusive and collectively exhaustive. Examples include occupation of Ghanaians (Farmer, Teacher, Doctor, Painter, Lawyer, etc.), gender (Male, Female), Ethnic group, Region of birth, brands of mobile phones, etc.
2. ***Ordinal*** – this is a qualitative, categorical scale that classifies data into distinct categories and also arranges them in a meaningful order or ranking. The ordinal scale provides information about the relative position or rank of each category. These categories have a clear order but the differences between them are not necessarily equal. Example: Ranks in GES (superintendent, Principal superintendent, assistant director II, etc.), model of mobile phones, Ranks in police service, Level of Education (Primary, JHS, SHS), Level of happiness (Very happy, Not happy, A little happy).
3. ***Interval*** – this quantitative scale orders data and also specifies the exact differences between values. It allows for meaningful comparisons of the magnitude of differences between items but does not have a true/absolute zero point. That is, recording a value of zero does not mean the absence of the trait or characteristic of interest. Examples are singing ability, score on a mathematics examination and temperature in Celsius. For the temperature, the difference between 20°C and 30°C is the same as the difference between 10°C and 20°C. However, 0°C does not represent the absolute absence of heat. etc.
4. ***Ratio*** – this is the highest level of quantitative measurement and it possesses all the characteristics of the interval scale but it also has a true/absolute zero point. Thus, recording a value of zero means an absence of the trait or characteristic of interest being measured. This allows for the full range of statistical operations, including meaningful comparisons of ratios between values. Ratios can be formed between values. Examples are money, age (in years), number of trees, number of tables, height in centimetres where 0 cm represents no height, and you can say someone is twice as tall as another person (e.g., 180 cm vs. 90 cm), etc.

Activity 9.8

1. Take a stroll around your campus/ house.
2. Write down the different data/variables/ characteristics of interest you identify.
3. Group them under nominal, ordinal, interval and ratio scales of measurement
4. Discuss your findings with a classmate.

Here is a summary of the hierarchy of data:

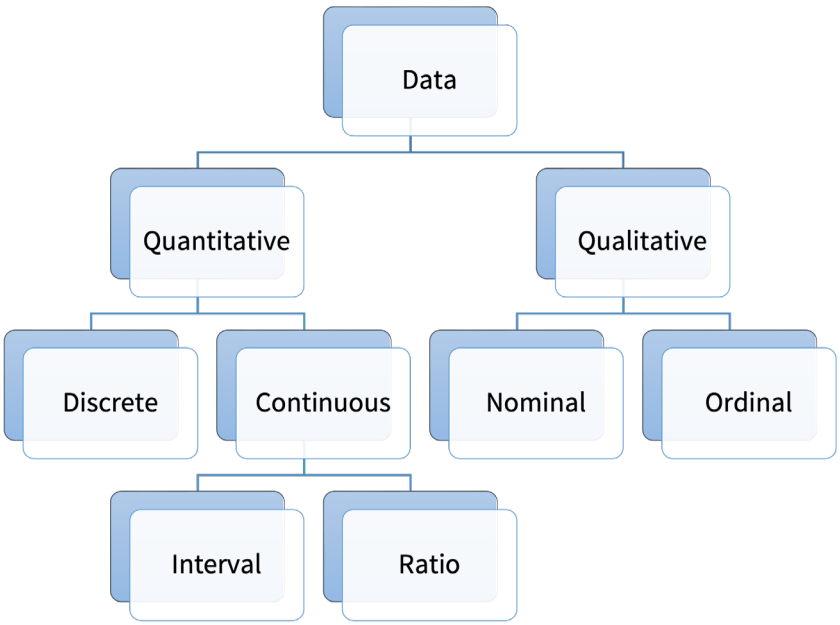


Figure 1: Types of data

TABULAR REPRESENTATION OF DATA- GROUPED & UNGROUPED DATA

Ungrouped data: this is the data you gather from an experiment that is not grouped into categories or classified.

For example, the ages of some randomly selected football players are as follows:

24, 23, 25, 23, 30, 24, 37, 25, 23, 22, 25, 22, 31, 29, 22, 25, 21, 25, 24, 24, 22.

This data is known as raw data as it has not been sorted into any categories.

Activity 9.9

1. Create a table made up of three columns: Age, Tally and Frequency.
2. Fill the age column with the distinct ages recorded in the example above. Do not repeat an age.
3. Now make a stroke every time you come across a particular age and record under the tally column.
4. Repeat this until all listed ages are exhausted
5. Count the number of strokes corresponding to each of the ages and record under the frequency column.

   The table has been started for you.

Table 2: Frequency Table

Age	Tally	Frequency
21	/	1
22	𝍸	...
....	...	...

This is an ungrouped tabular representation for the raw data. This is a much easier way to appreciate the ages.

Grouped data: this is data that has been categorised using **intervals**. It is usually used when the data is large and covers a large range. This makes it easier to spot patterns and trends.

To represent grouped data in a frequency distribution, we will need knowledge of class intervals, class boundaries and class mid-points.

Class intervals

The size of each class into which a range of variables is divided is the class interval.

For example, 10 – 14 (read as 10 to 14). This gives the number of values from 10 to 14 inclusive. 10 is the lower-class limit and 14 is the upper-class limit.

Consider the table below representing the outcome of research showing the aggregate of JHS graduates of Mamprobi Sempe JHS.

Table 3: Frequency Table

Aggregate	Number of students
6 – 10	3
11–15	8
16–20	22
21 –26	5

The class intervals are 6 – 10, 11 – 15, 16 – 20, 21 – 26. The smaller number in each class is called the lower-class limit, the bigger number is called the upper-class limit. This means 6, 11, 16 and 21 are lower class limits.

Class boundaries

To find the class boundary:

1. Subtract the upper-class limit of a class from the lower-class limit of the next immediate class.
2. Divide the result by two to obtain a number k.
3. Add k to all upper-class limits and subtract k from all lower-class limits and use inequality signs to remove any ambiguity about the class each value is in.

Activity 9.10

Using the data above, let us find the class boundaries of the data:

A reminder of the data of the aggregate of JHS graduates of Mamprobi Sempe JHS.

Table 4: Frequency Table

Aggregate	Number of students
6 – 10	3
11–15	8
15–20	22
21 –25	5

Solution

The first-class interval is 6−10, its lower-class limit is 6 and its upper-class limit is 10

The next immediate class has an interval of 11 − 15, its lower-class limit is 11 and the upper-class limit is 15.

Hence, using our steps.

Step 1. $11 - 10 = 1$

Step 2. $k = \frac{1}{2} = 0.5$

Step 3. Subtract k from the lower bound and add k to the upper bound of each class and use inequality signs to make it clear where each piece of data should go:

Table 5: Frequency Table

Class boundaries	Number of students
(6-0.5)−(10+0.5)	3
(11-0.5)−(15+0.5)	8
(16-0.5)−(20+0.5)	22
(21-0.5) −(25+0.5)	5

So, the finished table should look like this:

Table 6: Frequency Table

Class boundaries	Number of students
$5.5 \leq x < 10.5$	3
$10.5 \leq x < 15.5$	8
$15.5 \leq x < 20.5$	22
$20.5 \leq x < 25.5$	5

Class mid-point

This is defined as the mean of the lower and upper-class limits of a particular class interval.

Activity 9.11

The data below shows the mass (kg) of 40 students in a class. The measurement is to the nearest kg.

55	70	57	73	55	59	64	72
60	48	58	54	69	51	63	78
75	64	65	57	71	78	76	62
49	66	62	76	61	63	63	76
52	76	71	61	53	56	67	71

1. Create a table with four columns: Mass (kg), Class Boundaries, Frequency and Class midpoint.
2. Fill the Mass (kg) column with the intervals 45-49, 50-54, until all the masses are captured in a particular interval. Do not repeat an interval.
3. Count the number of observations that satisfy each interval and record under the frequency column.
4. Repeat until all observations are counted and recorded.
5. Add 45 to 49 and divide by 2. Record the result under the class midpoint column corresponding to the interval 45-49.
6. Repeat step 5 for all the intervals

The table has been started for you.

Table 7: Frequency Table

Mass (kg)	Class Boundaries	Frequency (f)	Class midpoint (x)
45 – 49	$44.5 \leq x < 49.5$	2	$\frac{45+49}{2} = \frac{94}{2} = 47$
50 – 54	$49.5 \leq x < 54.5$	4	$\frac{50+54}{2} = \frac{104}{2} = 52$
....		...	...

Example 1

The data shows the annual profits made by a sample of companies operating in Ghana.

Table 8: Frequency Table

Profit (Million$)	Number of companies
1– 4.5	8
4.6 –12.5	20
12.6 – 16.0	12
16.1 – 20.0	10

Find the class boundaries and the class mid-point.

Solution

$k = \frac{4.6 - 4.5}{2} = \frac{0.1}{2} = 0.05$

Table 9: Frequency Table

Class boundaries	Number of companies	Class mid-point
(1-0.05)–(4.5+0.05)	8	$\frac{1 + 4.5}{2} = \frac{1.5}{2} = 2.75$
(4.6-0.05)–(12.5+0.05)	20	$\frac{4.6 + 12.5}{2} = \frac{17.1}{2} = 8.55$
(12.6-0.05)– (16.0+0.05)	12	$\frac{12.6 + 16}{2} = \frac{28.6}{2} = 14.3$
(16.1-0.05)–(20.0+0.05)	10	$\frac{16.1 + 20}{2} = \frac{36.1}{2} = 18.05$

The table below shows the class boundaries and the mid-point:

Table 10: Frequency Table

Class boundaries	Number of companies	Class mid-point
$0.95 \leq x < 4.55$	8	2.75
$4.55 \leq x < 12.55$	20	8.55
$12.55 \leq x < 16.05$	12	14.3
$16.05 \leq x < 20.05$	10	18.05

GRAPHICAL REPRESENTATION OF CATEGORICAL DATA

Data can be represented in a table or graphically, both of which help us to see patterns, identify trends and make comparisons that might be hidden in raw data, but this tends to be truer for graphs. Choosing the right graph depends on the type of data and what is to be learnt from it. Graphs can be used in business reports and presentations, market research and consumer behaviour analysis, public policy and government reports, health care, medical data, etc.

Bar Charts or Bar Graphs

Bar charts are used to show numbers that are independent of each other. For example: ages of footballers, preference for a particular food, university, programme, etc. Bar charts are mostly used for comparisons. Each bar represents a category and its height shows the value. They work best for comparing things that don't change over time (e.g., student test scores in different subjects).

A **bar chart** uses rectangular bars to represent data. The bars can be horizontal or vertical.

We will look at two types of bar charts. The single bar chart and the double bar chart.

Activity 9.12

1. Read carefully the data from the table on students and their favourite fruits.

Table 11: Frequency Table

Favourite Fruit	Number of students
Watermelon	6
Banana	7
Apple	10
Mango	12
Orange	8

2. Draw two axes (one horizontal, one vertical).
3. Label the horizontal, "favourite fruit" and the vertical, "number of students".

4. Choose an appropriate scale for the number of students' axis and assign the numbers.
5. For each fruit, draw a bar that corresponds to the number of students who prefer that fruit.
6. Space the bars equally and indicate the names of the fruit for each bar.
7. Give the graph a title (e.g. "Favourite Fruits of Students").

Your graph should be similar to the one below.

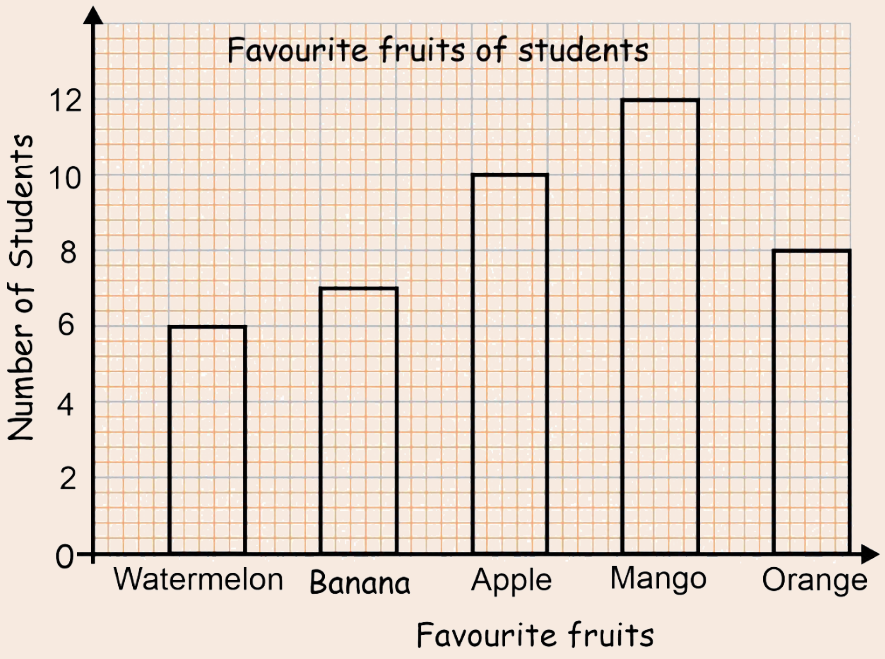


Figure 1: A bar graph on the favourite fruits of students.

Example 2

The data below shows the number of houses owned by a sample of Mokola women.

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

a) Draw a frequency distribution table for data.

b) Draw a bar chart for the data.

Solution

a)

Table 12: Frequency Table

Number of houses	Tally	Frequency
0	## /	6
1	## //	7
2	## ////	9
3	## /	6
4	////	4
5	## ///	8
Total		40

b) A bar chart showing number of houses owned by Makola women

Figure 2: A bar chart showing houses owned by Makola women

Example 3

The table below shows the rainfall distribution of two towns, Kasoa and Dansoman.

Table 13: Frequency Table

Month	Jan	Feb	March	April	May	June	July
Kasoa (mm)	6	10	12	14	17	13	16
Dansoman (mm)	8	12	10	16	17	14	17

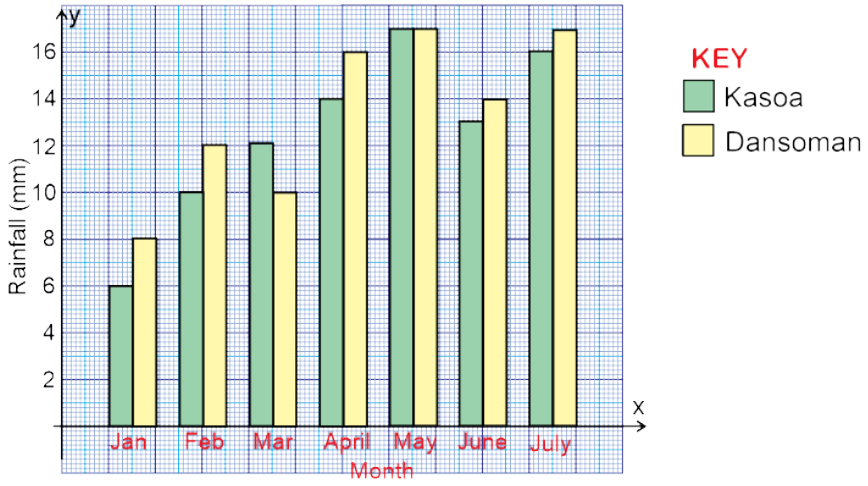
a) Draw a double bar chart for the data.

b) Which month was the wettest in Kasoa?

c) Which is the driest month in Dansoman?

Solution

a) Bar graph showing rainfall distribution in Kasoa and Dansoman



a) May was the wettest month in Kasoa.

b) January was the driest month in Dansoman.

Pie Charts

Pie Charts are used to show how a whole is divided into different parts. For example: how a budget is spent on items in a home within a month. They are mostly used to represent distinct or categorical data.

A pie chart displays data in the form of a circle. Each category of the data is represented by a sector.

$$\textbf{\textit{Angle of a sector of a category}} = \frac{\textit{value of the category}}{\textit{Total}} \times \mathbf{360°}$$

(recall that the sum of angles in a circle is **360°**)

Activity 9.13

1. The data in the table below gives students and their favourite pets.

Table 14: Frequency Table

Favourite Pet	Number of students
Dogs	12
Cats	6
Birds	4
Fish	8

2. Add up the total number of students.

3. Determine how much of the pie each category will take up.

For example, Angle for Dogs $= \frac{12}{30} \times 360° = 144°$

4. Remember to calculate the angle for each of the pets (categories).

5. Draw a circle.

6. Use a protractor to measure the angles for each category starting from a fixed point (e.g., the top of the circle).

7. Draw a line from the centre outwards to separate each of the measured angles.

8. Now, you have slices of a pie. Label each slice with the corresponding pet name and angle.

Your pie chart should look similar to the one below.

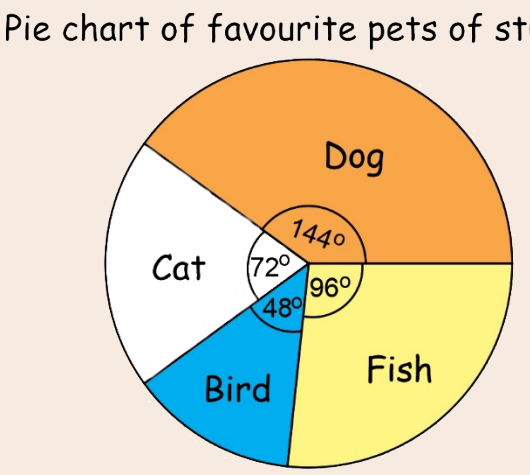


Figure 4: A pie chart showing the favourite pets of students

Example 4

The table below shows the favourite colours of learners in a class.

Table 15: Frequency Table

Colours	Number of learners
Red	12
Yellow	3
Green	x
Blue	4
White	8
Total	**30**

(a) Draw a pie chart to represent the data.

(b) Calculate the percentage of the class whose favourite colour is red.

Solution

(a) **Step 1:** Find the sector representing each category

Total number of learners $= 12 + 3 + x + 4 + 8 = 30$

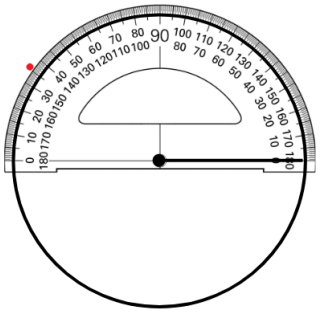
Green $= x = 30 - 12 - 3 - 4 - 8 = 3$

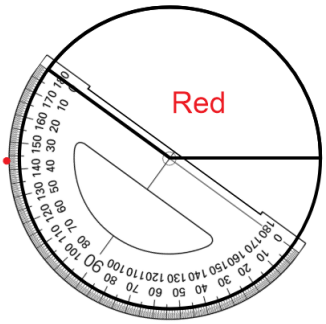
Table 16: Angle of sector calculation

Colours	Number of learners	Angle
Red	12	$\frac{12}{30} \times 360° = 144°$
Yellow	3	$\frac{3}{30} \times 360° = 36°$
Green	3	$\frac{3}{30} \times 360° = 36°$
Blue	4	$\frac{4}{30} \times 360° = 48°$
White	8	$\frac{8}{30} \times 360° = 96°$
Total	30	$360°$

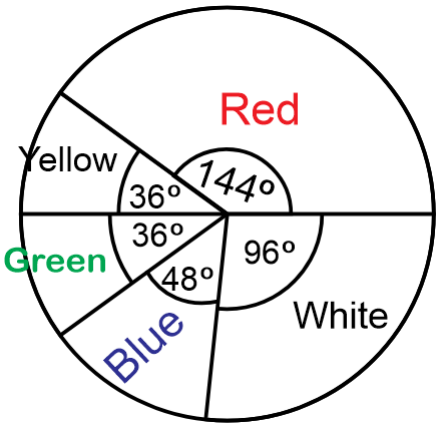
Step 2: With compasses construct a circle with a reasonable radius.

Step 3: Locate the centre and draw a radius. Use the protractor to measure the angles and draw the sectors. Label each sector with the colour and angle size.





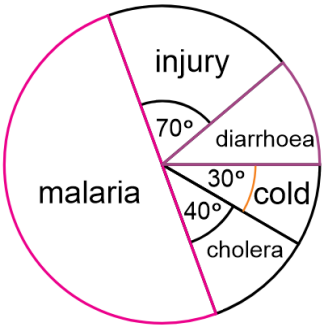
The final result will look like this



(b) Percentage of class whose favourite colour is red $= \frac{12}{30} \times 100\,\% = 40\%$

Example 5

The pie chart below shows the complaints a hospital received on a particular day.



Using the pie chart answer the following questions.

(a) Complete the table.

Table 17: Frequency Table

Complaint	Number of Patients
Diarrhoea	
Injury	
Malaria	
Cholera	
Cold	12
Total	

(b) Calculate the total number of patients who visited the hospital on that day.

(c) Find the percentage of patients who complained of malaria.

Solution

(a) Find the unknown angles and the corresponding number of patients:

Malaria occupies an angle on a straight line. So, malaria $=180°$

The sum of angles in a circle is $360°$.

$\therefore$ Angle for diarrhoea $= 360 - 180 - 70 - 30 - 40 = 40°$

From the pie chart, the angle representing Cold is $30°$ and this angle is equivalent to 12 patients.

i.e. $30° \equiv$ 12 patients

$360° \equiv$ Total number of patients

Total number of patients $= \frac{360°}{30°} \times 12 = 144$

The number of patients who complained of diarrhoea $= \frac{40°}{360°} \times 144 = 16$

The number of patients who were injured $= \frac{70°}{360°} \times 144 = 28$

The number of patients who complained of malaria $= \frac{180°}{360°} \times 144 = 72$

The number of patients who complained of cholera $= \frac{40°}{360°} \times 144 = 16$

Table 18: Frequency Table

Complaint	Number of Patients
Diarrhoea	16
Injury	28
Malaria	**72**
Cholera	**16**
Cold	12
Total	**144**

(b) Total number of patients = 36 + 54 + 216 + 144 + 90 = 144

(c) Percentage of patients who complained of malaria $= \frac{180}{360} \times 100\,\% = 50\%$

GRAPHICAL REPRESENTATION OF CONTINUOUS DATA

Graphical representation of continuous data is vital for visualising measurements like height or temperature within a range. Techniques include line graphs, histograms, and cumulative frequency curves. Understanding these helps interpret trends in science, economics, and decision-making.

Remember that continuous data is data that can take any value. Height, weight, temperature and length are all examples of continuous data.

Line Graphs

Line graphs are a way of displaying data that shows trends over time. Independent data values are usually placed on the horizontal axis (x-axis) and the dependent data values (sometimes frequency points) are placed on the vertical axis (y-axis). The dependent data values (frequency points) are connected using line segments. They are used to present numeric or quantitative data that occur over a time interval.

The main benefit of this type of graph is that you can spot **trends** in the data over time. This is a really useful tool in business to show sales trends or profits and losses over time or in science to show the growth rate of bacteria, change of pH or a change in temperature over time. Patterns in data allow for more accurate **predictions** in the future.

Types of line graphs

- A ***simple line graph*** is the most basic type of line graph. In this graph, only one dependent variable is tracked, so there is only a single line connecting all data points on the graph. All points on the graph relate to the same item, and the only purpose of the graph is to track the changes in that variable over time.
- In a ***multiple-line graph***, more than one dependent variable is charted on the graph and compared over a single independent variable (often time). Different dependent variables are often given different coloured lines to distinguish between each data set. Each line relates to only the points in its given data set.

Parts of a Line Graph

A good quality line graph will have most of the following characteristics:

Title – This is usually a simple statement written above the graph which explains what the graph is depicting.

Legend/Key – This explains what each dependent variable is and how to distinguish different sets of data. This feature is important when dealing with multiple and compound line graphs.

X-Axis – In most line graphs, this axis contains information that runs along the horizontal. It will be related to the independent variable or data. It should be clearly stated.

Y-Axis - The y-axis is the set of information that runs along the vertical axis. Some iterations of line graphs have this set of information on the right. These numbers count the items being measured. The graph may start at zero, though there are instances where it makes more sense to start at a higher number and the axis will clearly show it has been broken to allow this.

Line - Lastly, we draw the line. The line connects all data points within a single dependent variable. The line's movement shows the increase and decrease of information across time. It can also easily be compared against other lines as long as all data sets are being measured over similar periods.

Activity 9.14

1. Read carefully the data from the table on the temperature recorded each day of the week.

Table 19: Frequency Table

Day	Temperature (°C)
Monday	22
Tuesday	20
Wednesday	23
Thursday	25
Friday	24
Saturday	28
Sunday	26

2. Draw two axes (one horizontal, one vertical).
3. Label the horizontal, "Days" and the vertical, "Temperature ($°C$)".
4. Now, pick an appropriate scale for the Temperature axis and assign the numbers.
5. For each day, find the corresponding temperature and plot the point on the graph.

   For example, on Monday, plot a point at (Monday, 22°C), On Tuesday plot the point (Tuesday, 20°) etc.
6. After plotting all the points, connect the points with a straight edge.
7. Give the graph a title (e.g. "Temperature across the Week").

   Hopefully your graph looks like the one below.

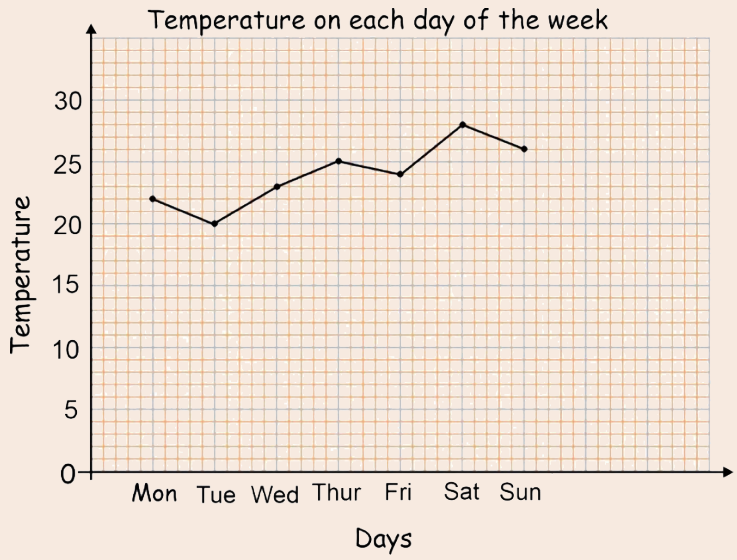


Figure 5: Line graph showing the temperature on each day of the week.

Example 6

The table below gives data on revenue and investment of Paintsiwa Automobile from 2015 to 2024. Values are in millions of Ghana Cedis.

Table 20: Data on revenue and investment

YEAR	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Revenue	1.00	2.00	2.40	2.00	3.00	2.40	2.00	3.60	3.40	3.60
Investment	1.20	1.40	1.40	2.00	2.40	2.00	2.40	1.60	2.40	2.20

(a) Present the data using a line graph

(b) In which year(s) did the company break even?

(c) In which year(s) did the company incur losses?

Solution

(a) A Multiple Line Graph showing Revenue and Investment of Paintsiwa Automobile.

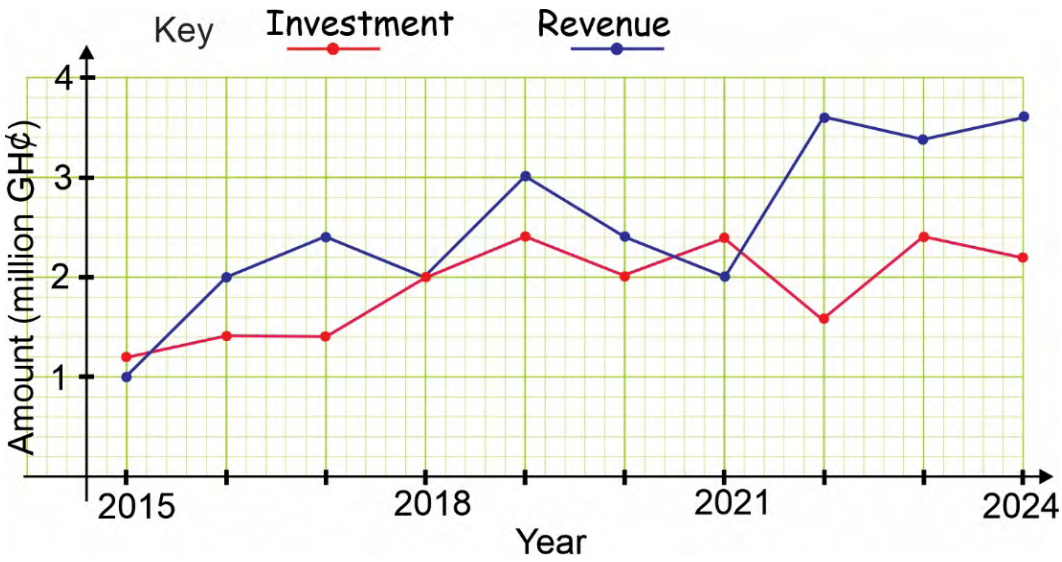


Figure 6: A compound line graph showing revenue and investment

(a) From the graph, the company broke even in 2018. This is the year the company made no profit as the investment equals the revenue.

(b) From the graph, the company incurred losses in 2015 and 2021. This is the year the Investment Points are **above** the Revenue points.

Example 7

The line graph below shows the expected and actual tomatoes harvested by a farmer from 2015 to 2024.

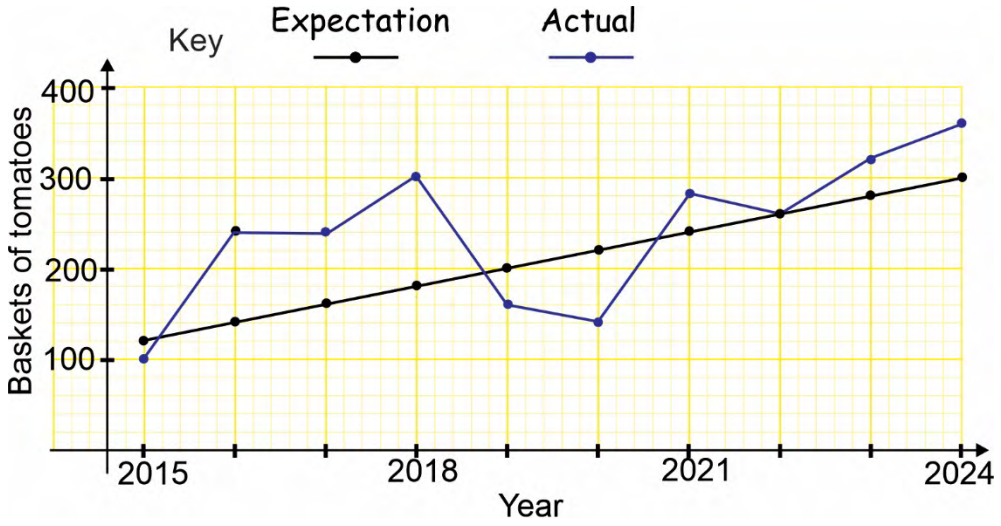


Figure 7: A compound line graph showing expected and actual tomatoes harvested

a) Copy and complete the table below

Table 21: Data on expected and actual tomatoes harvested

Year	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Expected	120					220				
Actual	100									

b) In which year did the farmer's expectations match reality?

c) In which years did the farmer harvest LESS than her expectations and by how many baskets?

d) Between which consecutive years did she have the highest increase in her actual harvest compared with the previous year.?

e) Compared with her expectations, in which year did she have the least harvest?

f) Compared with her expectations, in which year did she have the greatest harvest?

Solution

a)

Table 22: Complete data on expected and actual tomatoes harvested

Year	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Expected	120	140	160	180	200	220	240	260	280	300
Actual	100	240	240	300	160	140	280	260	320	360

b) On the graph, this is shown where the year two points intersect. Therefore, the farmer met her expectations in 2022.

c) This is the year(s) the blue points are below the black points. In other words, the year the actual points are below the expected points.

In 2015, less by 120−100=20 baskets

In 2019, less by 200−160=40 baskets

In 2020, less by 220-140=80 baskets

d) The greatest increase in actual results is illustrated on the graph with the steepest gradient. This occurs between 2020 and 2021.

e) To answer this question, we will first find the difference between the actual and expected harvest for each year.

Table 23: Data on difference between expected and actual tomatoes harvested

Year	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Expected	120	140	160	180	200	220	240	260	280	300
Actual	100	240	240	300	160	140	280	260	320	360
Difference	-20	100	80	120	40	**-80**	40	0	40	60

The year she harvested farthest below her expectation is 2020. She harvested 80 baskets less than expected.

f) Compared with her expectations, she had the greatest harvest in 2018. She exceeded her expectations by 120 baskets.

Histograms

The word histogram comes from the Greek word “histos”, which means pole or mast, and “gram”, which means chart or graph. Hence, the direct definition of “histogram” is “pole chart.” Perhaps this word was chosen because a histogram

looks like several side-by-side poles. It is defined as a diagram consisting of rectangles whose area is proportional to the frequency of a variable and whose width is equal to the class interval. It displays statistical information that uses rectangles to show the frequency of data in successive numerical intervals.

It consists of continuous (adjoining) boxes. It has both a horizontal axis and a vertical axis. Usually, the horizontal axis is labelled with what the data represents. This can be age, height, time, etc. The vertical axis represents the **frequency distribution.**

One advantage of the histogram is that it can display large data sets. It can also give you the shape, centre, and spread of the data

To construct a histogram, follow the following steps;

1. Divide the entire range of values into class intervals. The intervals are usually consecutive variables and do not overlap.
2. Count how many values fall into each interval.
3. If the intervals are of equal width, erect rectangles over the intervals with their height equal to the frequency of their respective classes (the number of values that fall into that class).

However, if the intervals are of **unequal** width, the height of each mounted rectangle is equal to the **frequency density** of that class.

Histograms are often confused with bar charts. A histogram is used for continuous data while the bar chart is a plot of categorical data. Before we solve examples remind yourselves about class width and boundaries we looked at earlier in this section.

Frequency Density

If f = frequency of a class and C = class width.

[Remember class width = Upper class boundary – Lower class boundary],

then, Frequency Density $= \frac{\textit{frequency of a class}}{\textit{class width of that class}} = \frac{f}{c}$

Example 8

Find the frequency density of the data in the table below.

Solution

Table 24: Frequency density table

Class boundaries	Class width (C)	*f*	Frequency density= *f*=
$0.95 \leq x < 4.55$	3.6	8	$\frac{8}{3.6} = 2.22$
$4.55 \leq x < 12.55$	8	20	$\frac{20}{8} = 2.5$
$12.55 \leq x < 16.05$	3.5	12	$\frac{12}{3.5} = 3.43$
$16.05 \leq x < 20.05$	4	10	$\frac{10}{4} = 2.5$

Activity 9. 15

1. Read carefully the data from the table on scores and the corresponding frequency.

Table 25: Frequency table

Scores	Frequency
40 – 49	3
50 – 59	7
60 – 69	6
70 – 79	5
80 – 89	5
90 – 99	4

2. Draw two axes (one horizontal, one vertical).
3. Label the horizontal axis, "scores" and the vertical axis, "frequency density".
4. Now, pick an appropriate scale for both axes and assign the numbers.

5. Add a column to the table to show the class boundaries (so there are no gaps). As the class widths are all equal there is no need to find the frequency density in this case.

Table 26: Frequency distribution

Scores	Class Boundaries	Frequency
40 – 49	$39.5 \leq x < 49.5$	3
50 – 59	$49.5 \leq x < 59.5$	7
60 – 69	$59.5 \leq x < 69.5$	6
70 – 79	$69.5 \leq x < 79.5$	5
80 – 89	$79.5 \leq x < 89.5$	5
90 – 99	$89.5 \leq x < 99.5$	4

6. For the class boundaries of each score range, draw a bar that corresponds to the frequency of that score. There must be no space between the bars.

7. Give the graph a title (e.g. "Examination scores of Students").

Your histogram should look similar to the one below.

Histogram of examination scores of students.

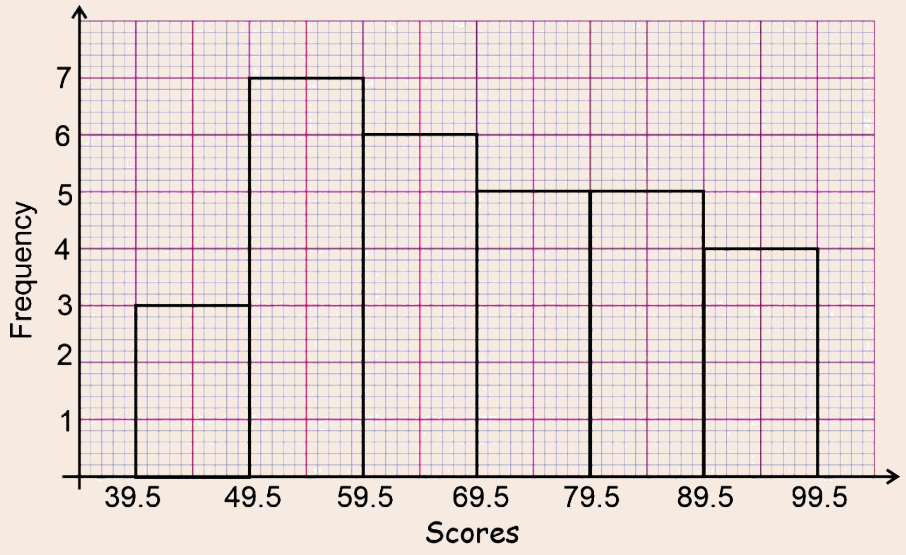


Figure 8: A histogram showing examination scores

Example 9

Dede researched the time taken, in minutes, by her classmates to get ready for school.

Her findings are shown below.

21	45	28	43	18	33	28	39	19	49
30	37	33	45	25	39	49	34	28	41
41	46	19	16	37	28	35	25	38	25
17	43	27	49	38	23	26	49	28	30
30	38	37	33	50	20	16	18	17	40
50	24	22	20	28	34	25	28	41	27

a) What is Dede's sample size?

b) Prepare a frequency distribution table using the groupings 16-20, 21-25, 26-30, etc

c) Draw a histogram to represent the data.

Solution

a) The sample size is the students from whom she collected the data. This is the same as the number of data values. By counting, we have 60. Therefore, Dede's sample size is 60.

b)

Table 27: Frequency distribution

Time (minutes)	Tally	Number of girls	Class boundaries
16 – 20	𝍸	10	$15.5 \leq x < 20.5$
21 – 25	𝍸 ///	8	$20.5 \leq x < 25.5$
26 – 30	𝍸 𝍸 ///	13	$25.5 \leq x < 30.5$
31 – 35	𝍸 /	6	$30.5 \leq x < 35.5$
36 – 40	𝍸 ////	9	$35.5 \leq x < 40.5$
41 – 45	𝍸 //	7	$40.5 \leq x < 45.5$
46 – 50	𝍸 //	7	$45.5 \leq x < 50.5$
Total		60	

c) As the class widths are all equal, there is no need to convert to frequency density.
Histogram showing the length of time taken to get ready for school.

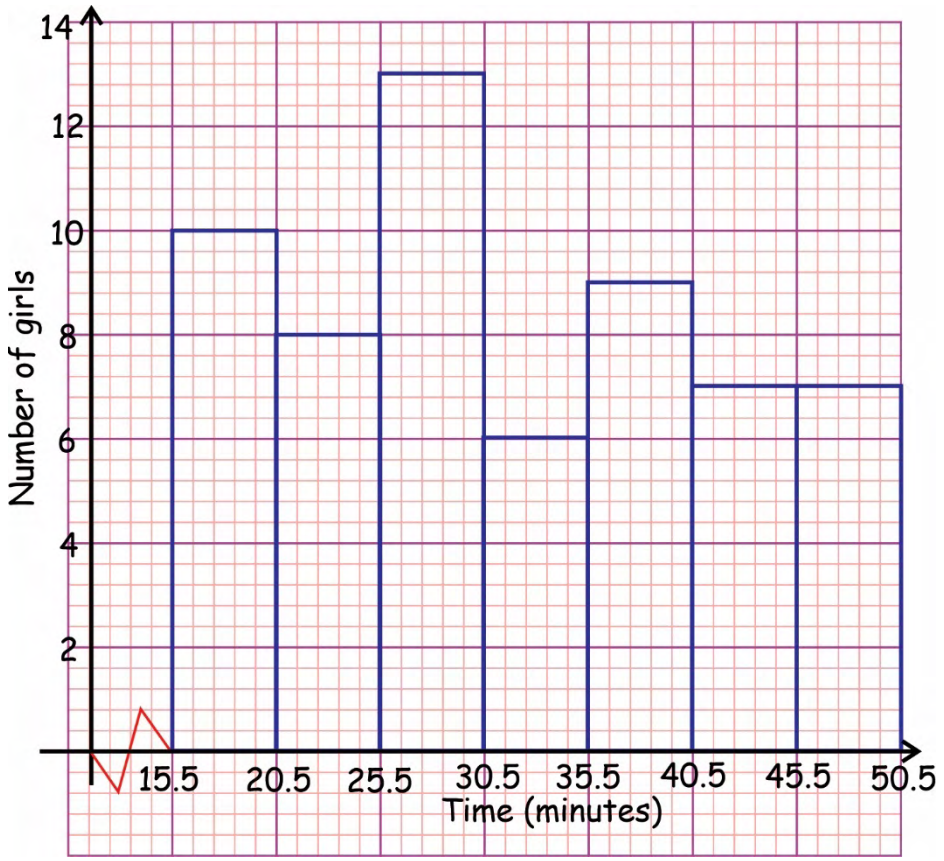


Figure 9: A histogram showing length of time it took to get ready for school

Example 10

Madam Wan is a technical support officer with Ghana Telecom. She took data on the duration of a sample of phone calls (minutes) she received from clients seeking support. The list is shown below.

5.4	4.6	28.3	13.4	13.5	25.0	33.0	3.2
27.9	9.5	34.0	19.0	10.9	30.3	7.1	17.7
0.7	23.1	12.2	13.7	18.6	31.7	23.8	11.5
6.4	16.2	24.3	32.4	5.0	25.7	20.6	26.6
9.0	8.3	14.7	1.4	16.8	33.2	3.5	35.4

(a) Construct a frequency distribution table using equal intervals of $0 \leq x < 6$, $6 \leq x < 12$, $12 \leq x < 18$, etc.

(b) Represent the frequency distribution table with a histogram.

Solution

(a)

Table 28: Frequency distribution

Duration(minutes)	Tally	frequency
$0 \leq x < 6$	𝍸 //	7
$6 \leq x < 12$	𝍸 //	7
$12 \leq x < 18$	𝍸 ///	8
$18 \leq x < 24$	𝍸	5
$24 \leq x < 30$	𝍸 /	6
$30 \leq x < 36$	𝍸 //	7

(b) Since the duration is already in boundaries, we proceed to the (b) part of the question.

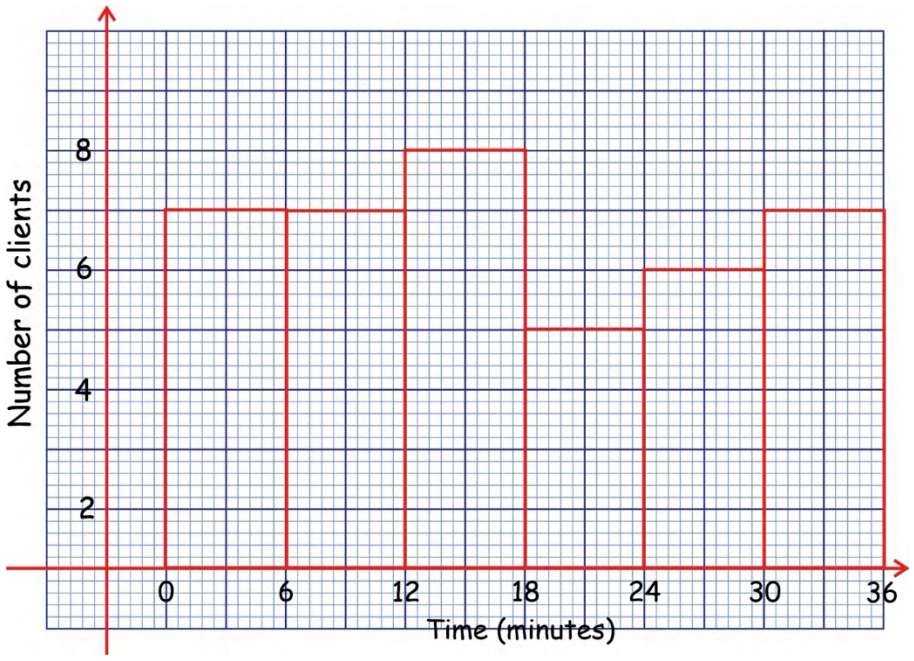


Figure 10: A histogram showing the duration of phone calls received

Example 11

The list below gives information about the weights (pounds) of a sample of women.

131	140	132	116	136	129	136	142
128	122	126	115	123	133	128	131
118	124	117	133	138	128	143	130
132	130	134	137	125	135	113	121
119	125	135	127	137	126	129	134
143	114	129	137	120			

(a) Copy and complete the frequency distribution table below

Table 29: Frequency distribution

Weight	Tally	Number of students
113 – 117		
118 – 127		
128 – 130		
131 – 137		
138 – 142		

(b) Draw a histogram for the data.

Solution

(a)

Table 30: Frequency distribution

Weight	Tally	Number of students
113 – 117	𝍸	5
118 – 127	𝍸𝍸 //	12
128 – 130	𝍸 ////	9
131 – 137	𝍸𝍸 ////	14
138 – 142	𝍸	5

(b) We first prepare a distribution table for the data.

Table 31: Frequency distribution

Weight	Tally	Number of students	Class Boundaries	Class width	Frequency Density= $\frac{f}{c}$
113 – 117	𝍸	5	$112.5 \leq x < 117.5$	5	1
118 – 127	𝍸 𝍸 //	12	$117.5 \leq x < 127.5$	10	1.2
128 – 130	𝍸 𝍸	9	$127.5 \leq x < 130.5$	3	3
131 – 137	𝍸 𝍸 ////	14	$130.5 \leq x < 137.5$	7	2
138 – 142	𝍸	5	$137.5 \leq x < 142.5$	5	1

Since the classes are of unequal width, we use the frequency density instead of the frequency to draw the histogram.

Below is the histogram representing weight of students.

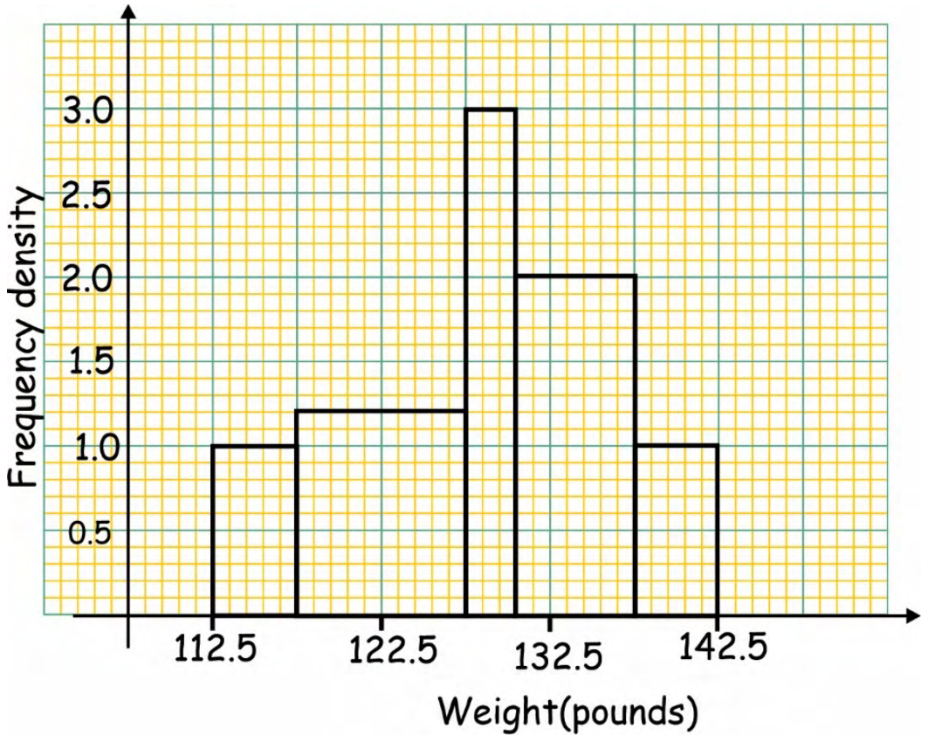


Figure 11: A histogram representing weights of students

Example 12

A policewoman records the speed of 792 vehicles on a road with a speed limit of 70km per hour. The histogram below represents the results.

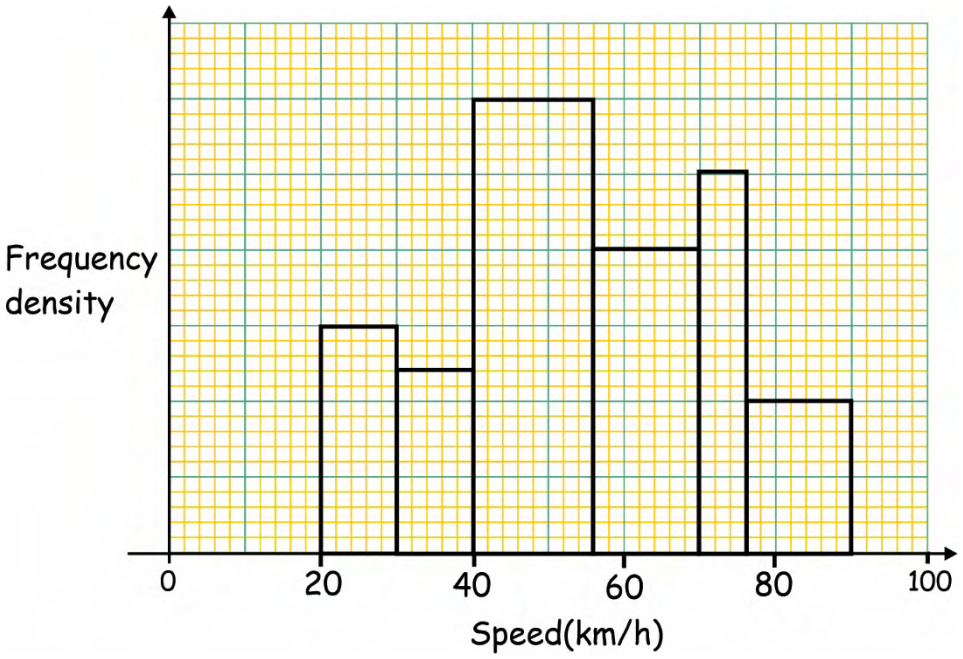


Figure 12: A histogram representing speed of vehicles without frequency density values

(a) Construct a frequency distribution table for the data.

(b) Calculate the number of vehicles that were travelling below the speed limit by at least 14km/h.

(c) Reconstruct the histogram to include the frequency density.

(d) If each driver who exceeded the speed limit was fined GH¢500.00. Find the total amount realized from the fines.

Solution

(a) The graph has not provided the frequency densities. However, we do have the total frequency and we can obtain the class boundaries from the graph.

Since the area of the rectangles is in proportion to the total frequency, we will apply ratio and proportion to find the frequency of each class.

To find the frequencies of each class interval, we will compare the area of each rectangle to the total frequency. To do this we will let the area of each small box be 1 square unit.

Starting from left,

This area of the first rectangle = 75

Square units since it contains 75 small boxes in it.

The area of the second rectangle = 60

The area of the third rectangle = 240

The area of the fourth rectangle = 140

Area of the fifth rectangle = 75

The area of the sixth rectangle = 70

The total area of the rectangles = 660

660 is proportional to 792 vehicles.

Frequency of the first rectangle $= \frac{75}{660} \times 792 = 90$

Frequency of the second rectangle$= \frac{60}{660} \times 792 = 72$

Frequency of the third rectangle$= \frac{240}{660} \times 729 = 288$

Frequency of the fourth rectangle$= \frac{140}{660} \times 729 = 168$

Frequency of the fifth rectangle= 90, same as the first rectangle.

Frequency of the last rectangle$= \frac{70}{660} \times 792 = 84$

∴ Frequency distribution table:

Table 32: Frequency distribution

Class Boundaries	Frequency/f	Class width/C	Frequency density
$20 \leq x < 30$	90	10	$\frac{F}{C} = \frac{90}{10} = 9$
$30 \leq x < 40$	72	10	7.2
$40 \leq x < 56$	288	16	18
$56 \leq x < 70$	168	14	12
$70 \leq x < 76$	90	6	15
$76 \leq x < 90$	84	14	6
	792		

(b) The speed limit is 70km/h. Travelling at least 14km/h below the speed limit is the same as travelling at a speed which is less than or equal to 70 − 14 = 56km/h

From the table, the number of vehicles travelling at a speed ≤ 56km/h = 90 + 72 + 288 = 450

(c)

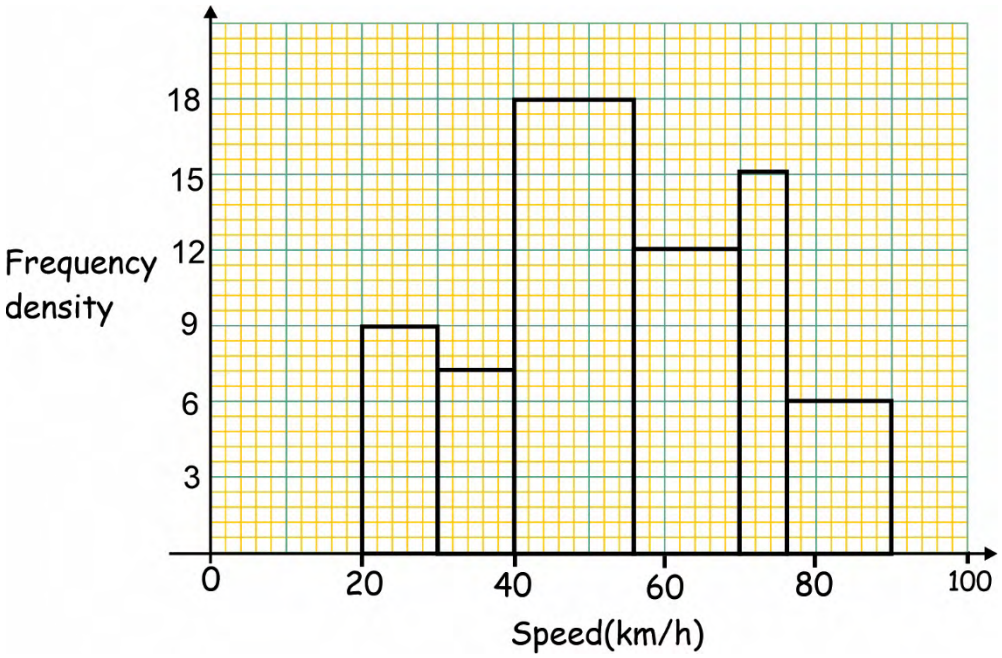


Figure 13: A histogram representing speed of vehicles with frequency density values

(d) From the table, vehicles which exceeded the speed limit of 70km/h = 90 + 84 = 174

If each driver is paying GH¢500.00 as a fine, the total amount = 174 × 500 = *GH¢* 87 000.00

Cumulative frequency Curves

Cumulative frequency is the running total of the frequency in a frequency distribution. A Cumulative Frequency curve is a graph plotted from a cumulative frequency table. A cumulative frequency curve is also called an ogive or cumulative frequency graph. A cumulative frequency curve can estimate a data set's median, quartiles and percentiles.

How to plot a Cumulative Frequency Curve:

Step 1. Draw and mark to scale the horizontal axis to allow for the range of data for the smallest class boundary to the largest.

Step 2. On the same graph sheet, draw and mark to scale the vertical axis with the accumulated frequencies.

Step 3. Plot the accumulated frequencies against respective **upper**-class boundaries.

Step 4. Join the points with a smooth free-hand curve.

Activity 9.16

1. The table below shows the ages of 30 people at a stadium.

Table 33: Frequency distribution

Age (years)	Frequency
10 – 19	4
20 – 29	6
30 – 39	8
40 – 49	5
50 – 59	4
60 – 69	3

2. Work out the class boundaries to show that the data is continuous and add up the frequencies to create the cumulative frequency column.

Table 33: Frequency distribution

Age (years)	Class Boundaries	Frequency	Cumulative Frequency
10 – 19	$9.5 \leq x < 19.5$	4	4
20 – 29	$19.5 \leq x < 29.5$	6	10
30 – 39	$29.5 \leq x < 39.5$	8	18
40 – 49	$39.5 \leq x < 49.5$	5	23
50 – 59	$49.5 \leq x < 59.5$	4	27
60 – 69	$59.5 \leq x < 69.5$	3	30

3. Draw two axes (one horizontal, one vertical).

4. Label the horizontal, "Age (years)" and the vertical, "Cumulative frequency".

5. Now, pick an appropriate scale for both axes and assign the numbers.

6. For each Age range, indicate the upper-class boundary, find the corresponding cumulative frequency and plot the point on the graph.

For example, for the interval 10-19, plot a point at (19.5, 4), and so on.

7. After plotting all the points, connect the points with a freehand curve.
8. Give the graph a title (e.g. "Ages of people in the stadium").

Example 13

Ama recorded the weights (kg) of a sample of luggage during check-in at Ghana International Airport. Her findings are listed below.

22	3	44	36	29	31	35	33	59	12
48	54	23	47	57	11	45	40	37	32
4	48	56	55	42	31	54	28	63	64
38	36	64	6	35	28	18	23	52	42
41	55	34	68	49	65	35	65	43	37
53	45	56	77	28	36	58	38	17	39
22	74	15	43	67	75	67	41	53	22
28	32	78	22	46	11	39	29	44	8

(a) Using class intervals of 0 – 9, 10 – 19, 20 – 29, etc. Construct a cumulative frequency distribution table for the data.

(b) Draw a cumulative frequency curve for the distribution.

Solution

(a) Cumulative frequency table

Table 34: Frequency distribution

Weight (kg)	Class Boundaries	Tally	Frequency	Upper class boundaries	Cumulative frequency
0 – 9	$0 \leq x < -9.5$	////	4	9.5	4
10 – 19	$9.5 \leq x < 19.5$	## /	6	19.5	6 + 4 = 10
20 – 29	$19.5 \leq x < 29.5$	## ## //	12	29.5	10 +12 = 22
30 – 39	$29.5 \leq x < 39.5$	## ## ## ///	18	39.5	22 + 18 = 40
40 – 49	$39.5 \leq x < 49.5$	## ## ## /	16	49.5	40 + 16 = 56
50 – 59	$49.5 \leq x < 59.5$	## ## //	12	59.5	56 + 12 = 68
60 – 69	$59.5 \leq x < 69.5$	## ///	8	69.5	68 + 8 = 76
70 – 79	$69.5 \leq x < 79.5$	////	4	79.5	76 + 4 = 80

(b) Let us use these steps to construct the cumulative frequency curve.

Step 1: Draw your axes and choose appropriate scales with the x axis representing weight (0 – 80 kg) and the y axis cumulative frequency (0 – 80).

Step 2: Plot the following ordered pairs on the graph (9.5, 4), (19.5, 10), (29.5, 22), (39.5, 40), (49.5, 56), (59.5, 68), (69.5, 76), (79.5, 80)

Step 3: Finally, join the points in step 3 with a single free-hand curve.

Your graph should be similar to the one below.

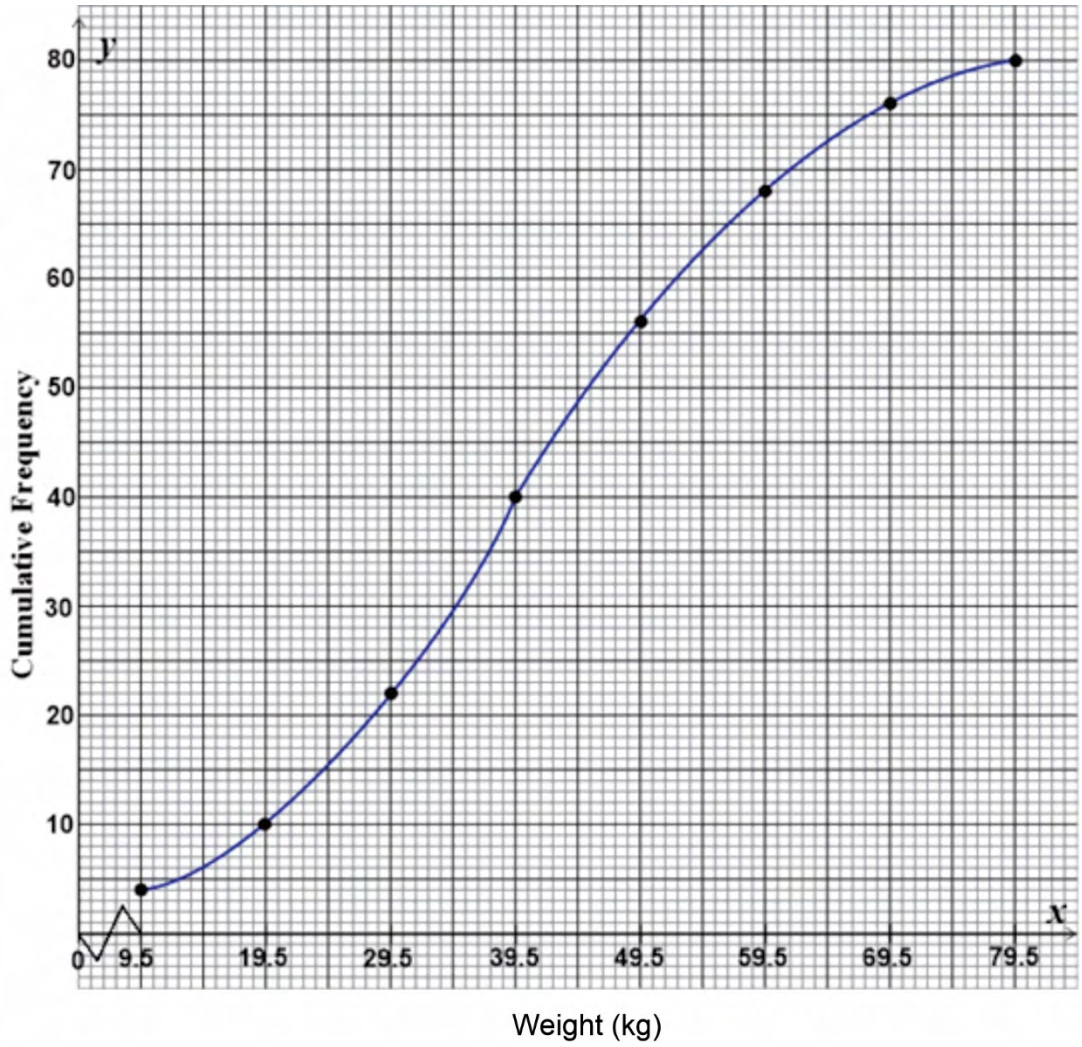


Figure 14: A cumulative frequency curve showing weights (kg) of luggage.

Example 14

The table below gives the frequency distribution of lengths(cm) of 40 bolts.

Table 35: Frequency distribution

Length of bolts(cm)	0.5 – 1	1.1 – 1.5	1.6 – 2	2.1 – 2.5	2.6 – 3	3.1 – 3.5	3.6 – 4
Number of bolts	1	2	4	8	16	7	2

(a) Prepare a cumulative frequency distribution table for the data

(b) Draw a cumulative frequency curve for the data

Solution

(a) We first construct a cumulative distribution table for the data

Table 36: Frequency distribution

Length of bolt(cm)	**Class Boundaries**	**Number of bolts**	**Upper-class boundary**	**Cumulative frequency**
0.5 – 1	$0 \leq x < 1.05$	1	1.05	1
1.1 – 1.5	$1.05 \leq x < 1.55$	2	1.55	3
1.6 – 2.0	$1.55 \leq x < 2.05$	4	2.05	7
2.1 – 2.5	$2.05 \leq x < 2.55$	8	2.55	15
2.6 – 3.0	$2.55 \leq x < 3.05$	16	3.05	31
3.1 – 3.5	$3.05 \leq x < 3.55$	7	3.55	38
3.6 – 4.0	$3.55 \leq x < 4.05$	2	4.05	40

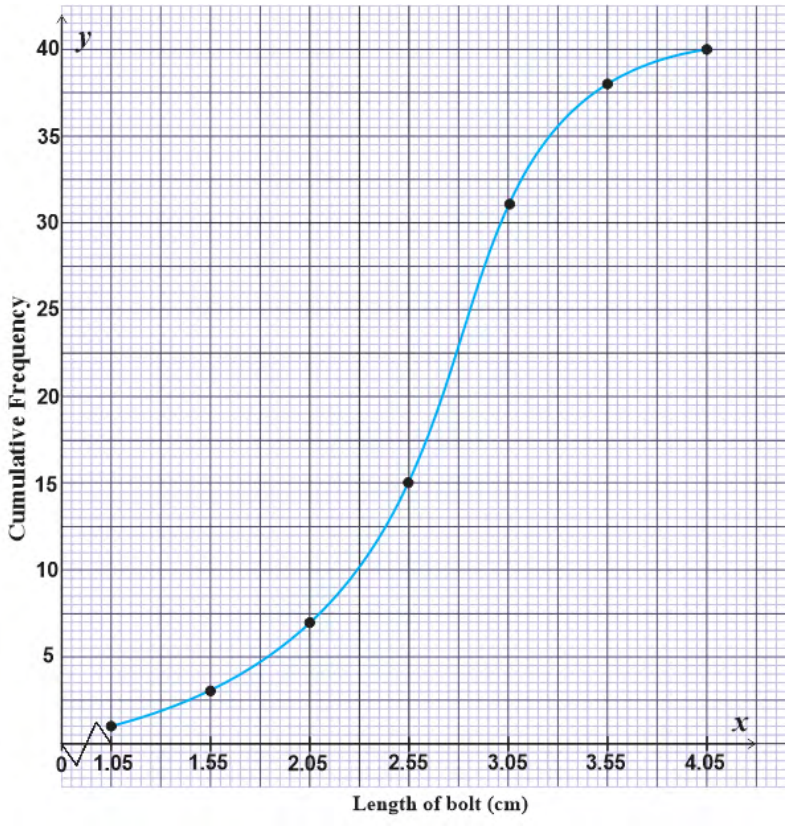


Figure 15: Cumulative frequency on length of bolt

Example 15

The graph below represents the profits of a sample of foreign companies.

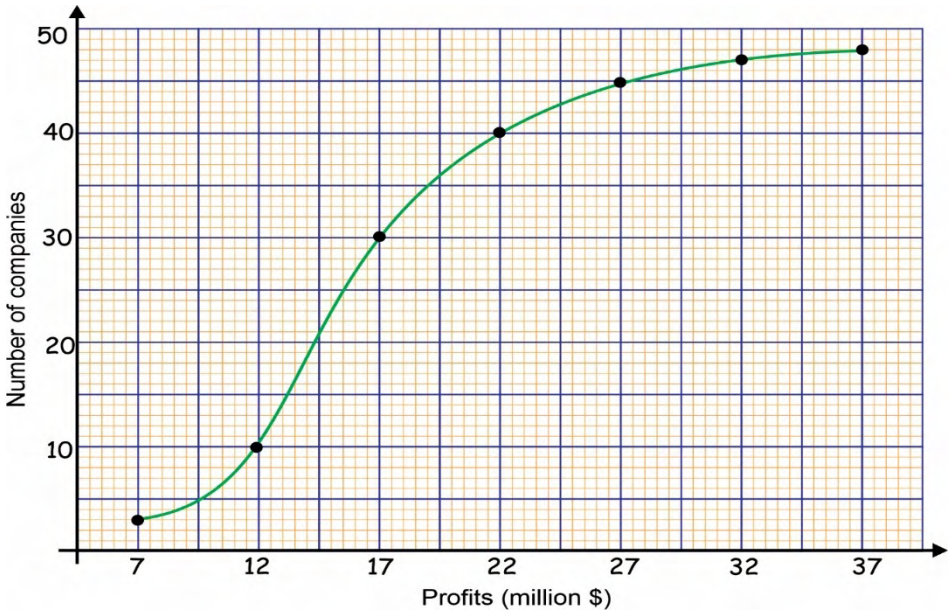


Figure 16: Cumulative frequency showing profits of a company

Use the graph to:

(a) Copy and complete the table below

Table 37: Frequency distribution

Class boundaries	Cumulative frequency	Frequency	

(b) Find the sample size

Solution

(a) From the graph, the upper-class boundaries are 7, 12, 17, 22, 27, 32 and 37

The class width = 12 − 7 = 5

Recall that; Class Width = Upper class boundary – Lower Class Boundary

This means the lower-class limit of the first class is 7 − 5 = 2

Therefore, the class boundary of the first class is 2 − 7 etc

The y-coordinates of the plotted points represent the cumulative frequency.

From the graph, the coordinates of the plotted points are (7, **3**), (12, **10**), (17, **30**), (22, **40**), (27, **45**), (32, **47**) and (37, **49**)

The frequency of the first class is the same as the cumulative frequency of the first class.

For the frequency of the second class, subtract the cumulative frequency of the first class from that of the second class. This gives 10 − 3 = 7. This pattern continues.

Therefore, the completed table looks like this:

Table 38: Frequency distribution

Class boundaries (millions of $)	Cumulative frequency	Frequency
$2 \leq x < -7$	3	3
$7 \leq x < 12$	10	10 − 3 = 7
$12 \leq x < 17$	30	30 − 10 = 20
$17 \leq x < 22$	40	40 − 30 = 10

Class boundaries (millions of $)	Cumulative frequency	Frequency
$22 \leq x < 27$	45	$45 - 40 = 5$
$27 \leq x < 32$	47	$47 - 45 = 2$
$32 \leq x < 37$	48	$48 - 47 = 1$
Total	48	48

(b) The sample size is the same as the total number of companies.
Therefore, the sample size is 48

Justifying the choice of a particular representation

Given the numerous graphical representation techniques discussed, we need to be able to determine the most appropriate one to use at any point in time. For instance, if you need to represent the number of goals scored by each football team in the Ghana premier league, will you use a bar chart, pie chart, line (time series) graph or a cumulative frequency curve? To justify your choice, we need to consider the data type, level of measurement and purpose of visualisation.

a. ***Type of data*** – consider whether the data is categorical or numerical. For categorical data with no ranking, use of bar charts and pie charts are most appropriate. For numerical data, histograms and line graphs (for time trends) work best.

b. ***Level of measurement*** – the nominal, ordinal, interval or ratio nature of the data also influences the choice of representation. Bar charts and pie charts are appropriate for nominal data. Line graphs, histograms and cumulative frequency curves are most appropriate for data from ratio and interval scales of measurement.

c. ***Purpose of data visualisation*** – we consider what we would like to attain through the representation. For instance, if it is to make comparison, bar charts are appropriate, if it to measure trends over time, line graphs work best, for proportions, pie charts do a great job too.

Note: *You need to consider the data type, level of measurement and purpose of visualisation together before making a choice of graphical representation. It is not an either-or situation but a combination of all.*

MEASURES OF CENTRAL TENDENCY

Central tendency (averages) describes how data points cluster around a middle value or the centre. There are five averages. Among them, the arithmetic mean, median and mode are called simple averages and the other two averages geometric mean and harmonic mean are called special averages. These averages are a way of describing the centre of data. We will focus on the simple averages.

Mode (ungrouped and grouped data)

The most frequent value in the data. Imagine a "fashion" - the mode is the most popular choice. In social science, the mode can be used to determine the most common response or opinion in a survey.

To determine the mode of a given raw data, identify the value(s) with the highest frequency or the most occurring value.

Mode from a grouped distribution

Mode is the item with the highest frequency. To calculate the mode from a grouped frequency distribution, use the following formula:

$$Mode = L + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right]c$$

Where L = Lower class boundary of the modal class

Δ_1 = excess frequency of modal class over the frequency of the next lower class.

Δ_2 = excess frequency of modal class over the frequency of the next higher class.

c = class size of modal class

Example 16

The data below shows the outcome of a survey on family sizes of a sample of residents at Ejisu.

Table 39: Frequency distribution

Family size	1 – 3	4 – 6	7 – 9	10 – 12	13 – 15
Number of families	8	23	12	3	1

Calculate an estimate for the mode and interpret your answer.

Solution

The class with the highest frequency is $4 - 6$

The class boundary of this class is $3.5 - 6.5$

The class size of the modal class (C) $= 6.5 - 3.5 = 3$

Lower class boundary of the modal class (L)=3.5

Excess frequency of modal class over the frequency of the next lower class (Δ_1) $= 23 - 8$

Excess frequency of modal class over the frequency of the next higher class (Δ_2) $= 23 - 12$

$$Mode = L + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right]c$$

$$Mode = 3.5 + 3 \times \left[\frac{23 - 8}{(23 - 8) + (23 - 12)}\right]$$

$$Mode = 3.5 + 3\left(\frac{15}{15 + 11}\right)$$

$$Mode = 4.077$$

Interpretation: An estimation of the family size of most of the sampled families in Ejisu is about 4.

Median (ungrouped and grouped data)

The middle value when the data is ordered from least to greatest. Think of a balanced seesaw - the median is the "centre point", or the fulcrum. The median income is used in economics to represent the typical income of a population, as it is less influenced by extremely high or low incomes.

The median is the midpoint of the data set. At this point, half the data values are above the value and half are below the value. This means that in a given data if the median is, for example, 40, it means half of the values are below (less than) or equal to 40 and half of the data values are above (greater than) or equal to 40.

To calculate the median for ungrouped data;

- Arrange the data in ascending or descending order and locate the middle value.

- If the number of observations (n) is odd, the median is the middle value of the ordered data.
- If n is even, the median is the average of the two middle values.
- For odd n, the median is the value at position $\frac{n+1}{2}$
- For even n, the median is the average of the values at positions $\frac{n}{2}$ and $\frac{n}{2}+1$.

Calculating the median from grouped data

Use the formula:

$$Median = L_1 + \left[\frac{\frac{N}{2} - cf_l}{f_m}\right]c$$

Where L_1 = lower-class boundary of the median class

N = total frequency

cf_l = cumulative frequency of classes lower than the median class

f_m = frequency of the median class

c = class size of median class

Example 17

The table below gives information on the observed lifetime (days) of solar bulbs.

Table 40: Frequency distribution

Lifetime(days)	100-119	120-139	140-159	160-179	180-199	200-219
Number of bulbs	62	78	75	50	105	80

Calculate an estimate for the median and interpret your answer.

Solution

Total frequency (n)= 62 + 78 + 75 + 50 + 105 + 80 = 450

So, the median class will be the class containing the 225th and 226th bulbs.

Table 41: Frequency distribution

Lifetime(days)	100-119	120-139	140-159	**160-179**	180-199	200-219
Number of bulbs	62	78	75	**50**	105	80
Cumulative freq	62	140	215	**265**	370	450

This means that the median class is 160 – 179.

The median class boundary is 159.5 – 179.5

Class size of median class (c) = 179.5 – 159.5 = 20

Lower class boundary of median class (L_1) is 159.5

Frequency of median class (f_m) = 50

Cumulative frequency of the preceding class (cf_l) = 215

$$Median = L_1 + \left[\frac{\frac{N}{2} - cf_l}{f_m}\right]c$$

$$Median = 159.5 + 20 \times \left[\frac{\frac{450}{2} - 215}{50}\right]$$

$$= 159.5 + 20\left(\frac{10}{50}\right) = 159.5 + 4 = 163.5$$

Interpretation: Half of the observed light bulbs have life days less than or equal to 163.5 and the other half have life days greater than or equal to 163.5. In other words, 225 of the observed bulbs have life days less than or equal to 163.5 and the other 225 bulbs have life days greater than or equal to 163.5 days.

Mean (ungrouped and grouped data)

The mean (often called the average) of all the data values is calculated by adding all values and dividing by the number of data points. The mean is used in financial analysis to calculate average returns on investments.

Arithmetic mean or mean

It is defined as the sum of the observations divided by the number of observations. It is denoted by the symbol $\bar{x}$ (x with a bar over it, pronounced "x bar"). The arithmetic mean of data is the average of all numbers. To determine the mean of a given raw data, sum all values and divide by the total number of observations (n).

If $x_1, x_2, x_3, \ldots\ldots\ldots\ldots, x_n$ are n numbers, the mean $(\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$.

If the data is in a frequency table then the mean $= \frac{\sum fx}{\sum f}$, where x represents the values and f the frequency of their occurrence.

Example 18

Determine the mean of the numbers 3, 8, 2, 10, 18, 9, 13.

Solution

$x_1 = 3$, the first number in the list

$x_2 = 8$, the second number on the list

$x_3 = 2$, the third number on the list

$x_4 = 10$,

$x_5 = 18$

$x_6 = 9$

$x_7 = 13$, 7^{th} number on the list.

Therefore,

$$\sum_{i=1}^{7} x_i = \textit{sum of numbers} = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7$$

$$= 3 + 8 + 2 + 10 + 18 + 9 + 13$$

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{3 + 8 + 2 + 10 + 18 + 9 + 13}{7} = \frac{63}{7} = \mathbf{9}$$

Example 19

Calculate the mean of the following set of numbers

11, 19, 12, −10, 19, 17, −8, 9, 27, −6

Solution

$$\text{Mean} = \frac{11 + 19 + 12 + (-10) + 19 + 17 + (-8) + 9 + 27 + (-6)}{10} = \frac{90}{10} = 9$$

Example 20

Without using a calculator, find the mean of the set of numbers:

$2\frac{2}{5}, -2\frac{1}{3}, 1\frac{1}{2}, -\frac{7}{4}$ and $2\frac{3}{5}$

Solution

$$\text{Mean} = \left(2\tfrac{2}{5} - 2\tfrac{1}{3} + 1\tfrac{1}{2} - \tfrac{7}{6} + 2\tfrac{3}{5}\right) \div 5$$

$$= \left(\tfrac{12}{5} - \tfrac{7}{3} + \tfrac{3}{2} - \tfrac{7}{6} + \tfrac{13}{5}\right) \div 5$$

$$= \left(\frac{(6 \times 12) - (10 \times 7) + (15 \times 3) - (5 \times 7) + (6 \times 13)}{30}\right) \div 5$$

$$= \left(\frac{72 - 70 + 45 - 35 + 78}{30}\right) \div 5$$

$$= \frac{90}{30} \times \frac{1}{5}$$

$$= \frac{3}{5}$$

Example 21

Calculate the mean of the set of numbers. Leave your answer in surd form

$7, \sqrt{5}, -\sqrt{45}, -13, 2\sqrt{125}, 2\sqrt{5}, 3\sqrt{20}$ and -10

Solution

$$\text{Mean} = \frac{7 + \sqrt{5} - \sqrt{45} + (-13) + 2\sqrt{125} + 2\sqrt{5} + 3\sqrt{20} + (-10)}{8}$$

$$= \frac{-16 + \sqrt{5} - \sqrt{9 \times 5} + 2\sqrt{25 \times 5} + 2\sqrt{5} + 3\sqrt{4 \times 5}}{8}$$

$$= \frac{-16 + \sqrt{5} - 3\sqrt{5} + 10\sqrt{5} + 2\sqrt{5} + 6\sqrt{5}}{8}$$

$$= \frac{-16 + 16\sqrt{5}}{8}$$

$$= \frac{16(-1 + \sqrt{5})}{8}$$

$$= \mathbf{2(-1 + \sqrt{5})}$$

Example 22

The table below shows the number of papers published by a sample of 25 researchers.

Table 42: Frequency distribution

Number of papers	4	5	6	7	8
Number of researchers	1	8	10	2	4

Calculate the mean and interpret your answer.

Solution

Table 43: Frequency distribution

Number of papers	4	5	6	7	8
Number of writers	1	8	10	2	4
Total number of papers	$4 \times 1 = 4$	$5 \times 8 = 40$	$6 \times 10 = 60$	$7 \times 2 = 14$	$8 \times 4 = 32$

$$\text{Mean} = \frac{(4 \times 1 + 5 \times 8 + 6 \times 10 + 7 \times 2 + 8 \times 4)}{25}$$

$$= \frac{4 + 40 + 60 + 14 + 32}{25}$$

$$= \frac{150}{25} = 6 \textbf{ papers}$$

Interpretation:On average each researcher published 6 papers.

Example 23

The list shows the aggregates of a sample of BECE graduates of UG basic school.

7	6	7	7	6	6	9	10	7	9
8	9	6	8	6	7	8	6	8	6
6	8	9	7	8	6	7	9	6	8

(a) Make a frequency distribution table for the data

(b) Use your distribution table to calculate the mean grade and correct your answer to the nearest whole number. Interpret your answer.

Solution

(a) Table 44: Frequency distribution table

Grades (x)	Tally	Frequency (f)	fx
6	𝍸 𝍸	10	$6 \times 10 = 60$
7	𝍸 //	7	$7 \times 7 = 49$
8	𝍸 //	7	$8 \times 7 = 56$
9	𝍸	5	$9 \times 5 = 45$
10	/	1	$10 \times 1 = 10$
		$\sum f = 30$	$\sum fx = 240$

(b) Mean grade $= \frac{\sum fx}{\sum f} = \frac{220}{30} = 7\frac{1}{3} \approx 7$ (to the nearest whole number)

Interpretation: On average, each graduate of UG basic School had an aggregate of 7.

Assumed method of calculating the mean

Using the assumed mean method, the mean is obtained using the formula

$$Mean = A + \frac{\sum fd}{\sum f} \text{ or } Mean = A + \sum \frac{d}{n}$$

Where A=Assumed mean

x = class mark

$d = deviation = x - A$

f = *the* frequency of individual classes

A = Assumed mean (usually any of the x values)

fx = frequency of individual classes multiplied by their corresponding class mark

$\sum$ = sum of,

$\sum fd$ = sum of fd,

$\sum f$ =sum of frequencies,

$\sum d$ = sum of deviations

n = total frequency or sum of frequencies

Example 24

The list below shows the number of houses owned by 8 realtors.

6, 4, 11, 8, 5, 4, 10, 8

Use the assumed mean method to calculate the mean and interpret your answer.

Solution

The class marks (x) are 6, 4, 11, 8, 5, 4, 10, 8.

We can choose any of these numbers as our assumed mean.

Let us choose 6 as our assumed mean. That is A = 6

6, 4, 11, 8, 5, 4, 10, 8

This means our deviations ($d = x - A$) will be

$6 - 6, 4 - 6, 8 - 6, 11 - 6, 5 - 6, 4 - 6 , 10 - 6, 8 - 6$

$d \rightarrow 0, -2, 2, 5, -1 , -2 , 4, 2$

$\sum d = 0 + (-2) + 2 + 5 + (-1) + (-2) + 4 + 2 = 8$

Mean= $A + \frac{\sum d}{n}$

A= 6, n = 8 since there are a total of 8 realtors.

Mean= $6 + \frac{8}{8} = 6 + 1 = 7$ houses.

We could have used any of the class marks as the assumed mean and we will arrive at the same answer. For example, let use A = 10 as the assumed mean this time.

$6 - 10, 4 - 10, 8 - 10, 11 - 10, 5 - 10, 4 - 10 , 10 - 10, 8 - 10$

$d \rightarrow -4, -6, -2, 1, -5, -6, 0, -2$

$\sum d = -4 + (-6) + (-2) + 1 + (-5) + (-6) + 0 + (-2) = -24$

Mean = $10 + \frac{-24}{8} = 10 - 3 = 7$ houses.

Interpretation: On average each of the eight realtors owns 7 houses.

Example 25

The table shows the number of days a sample of students were absent from school in a term

Table 45: Frequency distribution

Number of days absent	2	3	4	5	6	7
Number of students	5	8	6	7	1	3

Using an assumed mean of 5, calculate the mean.

Solution

Assume that A = 5.

$d = x - A = x - 5$

Table 46: Frequency distribution

Number of days absent (x)	Number of students (f)	$d = x - 5$	fd
2	5	$2 - 5 = -3$	$5 \times -3 = -15$
3	8	$3 - 5 = -2$	$8 \times -2 = -16$
4	6	$4 - 5 = -1$	$6 \times -1 = -6$
5	7	$5 - 5 = 0$	$7 \times 0 = 0$
6	1	$6 - 5 = 1$	$1 \times 1 = 1$
7	3	$7 - 5 = 2$	$3 \times 2 = 6$
	$\sum f = 30$		$\sum fd = -30$

$$Mean = A + \frac{\sum fd}{\sum f} = 5 + -\frac{30}{30} = 5 - 1 = 4$$

Therefore, the mean is 4. On average each student missed 4 days of school.

Example 26

The mean of 18, 13, 10, $2m$, $m + 1$ and 15 is 13, find the value of m

Solution

$$\text{Mean} = \frac{\sum x}{n}$$

$$13 = \frac{18 + 13 + 10 + 2m + (m + 1) + 15}{6}$$

$$13 \times 6 = 3m + 57$$

$$78 = 3m + 57$$

$$78 - 57 = 3m$$

$$21 = 3m$$

$$7 = \boldsymbol{m}$$

Finding the mean of grouped distributions

The procedure is similar to that of ungrouped data. The only difference is that, with a grouped data, replace the class marks with the class midpoint of each class interval.

Mean $= \frac{\sum fx}{\sum f}$, where x represents the mid-point of the class and f the frequency of the occurrence. Note that this will give you an *estimate* of the mean as we are making the assumption that all the data point in the class will 'average out' to be in the centre of the class.

Remember that midpoint $= \frac{upper\ class\ limit + lower\ class\ limit}{2}$

Example 27

The frequency distribution below shows the weight (kg) of a sample of animals at Accra Zoo.

Table 47: Frequency distribution

Weight(kg)	7–11	12–16	17–21	22–26	27–31	32–36	37–41	42–46
Number of animals	1	4	6	10	12	12	3	2

Calculate an estimate of the mean weight of the sampled animals.

Solution

Mean $= \frac{\sum fx}{\sum f}$, where $x = midpoint$

Table 48: Frequency distribution

Weight(kg)	*f*	Midpoint *x*	*fx*
7–11	1	$(7 + 11) \div 2 = 9$	$1 \times 9 = 9$
12–16	4	$(12 + 16) \div 2 = 14$	$4 \times 14 = 56$
17–21	6	$(17 + 21) \div 2 = 19$	$6 \times 19 = 114$
22–26	10	$(22 + 26) \div 2 = 24$	$10 \times 24 = 240$
27–31	10	$(27 + 31) \div 2 = 29$	$12 \times 29 = 290$
32–36	12	$(32 + 36) \div 2 = 34$	$12 \times 34 = 408$
37–41	5	$(37 + 41) \div 2 = 39$	$3 \times 39 = 195$
42–46	2	$(42 + 46) \div 2 = 44$	$2 \times 44 = 88$
	$\sum f = 50$		$\sum fx = 1400$

Mean $= \frac{1400}{50} = 28$

The estimated mean weight of the sampled animals is **28kg.**

Now, let us solve the same question using the 'assumed mean' method. We will choose one of the midpoints as our assumed mean. Let us go with 24.

This means A = 24.

Therefore, $d = x - 24$

Table 49: Frequency distribution

Weight(kg)	*f*	*x*	$d = x - 24$	*fd*
7–11	1	9	$9 - 24 = -15$	$1 \times -15 = -15$
12–16	4	14	$14 - 24 = -10$	$4 \times -10 = -40$
17–21	6	19	$19 - 24 = -5$	$6 \times -5 = -30$
22–26	10	24	$24 - 24 = 0$	$10 \times 0 = 0$
27–31	10	29	$29 - 24 = 5$	$10 \times 5 = 50$
32–36	12	34	$34 - 24 = 10$	$12 \times 10 = 120$
37–41	5	39	$39 - 24 = 15$	$5 \times 15 = 75$
42–46	2	44	$44 - 24 = 20$	$2 \times 20 = 40$
	$\sum f = 50$			$\sum fd = 200$

$$\text{Mean} = 24 + \frac{200}{50} = 24 + 4 = 28$$

On average, the estimated weight of the sampled animals at Accra Zoo is 28kg

Properties of the mean

1. If a constant number r is added to each of the data values, then the new mean will be $\bar{x} + r$
2. If a constant number r is subtracted from each of the data values, then the new mean will be $\bar{x} - r$
3. If a constant number is multiplied by each of the data values, the new men will be $\bar{x} \times r = r\bar{x}$
4. If each of the data values are divided by a constant non-zero number, r, then the new mean is $\frac{\bar{x}}{r}$
5. The sum of all deviations from the mean is 0.
6. The mean uses all values in the data set
7. The mean is affected by outliers.

Effect of extreme values on a measure of central tendency

A measure of central tendency should fairly describe a given dataset. We need to know if a particular measure of central tendency is appropriate to use when reporting findings on a given dataset.

To look into this we need to know more about the following terms:

1. **Outliers:** Outliers in statistics mean values in data that are unusual compared with the rest of the data by either being exceptionally large or exceptionally small.

Example 28

Below is the list of marks of students in a test:

35, 39, 35, 85, 40, 39, 37, 40, 90, 47, 46 and 45. Determine the outliers in this dataset.

Solution

The marks of most of the students were between 35 and 47. However, the scores of two learners are far higher compared to the rest of the students. Their marks were

90 and 85. The values 85 and 90 are called outliers since they are exceptionally large compared with the rest of the values.

Example 29

The data below show the weights (kg) of a sample of luggage:
98, 82, 90, 102, 105, 12, 92, 75, 8, 88, 86. Determine the outliers in this dataset.

Solution

The outliers in this data are 12 and 8 since these weights are exceptionally low compared with those of the other luggage.

2. **Symmetric data:** Symmetric data is where the mean, median and mode have the same value. One side of the distribution is a mirror image of the other side.

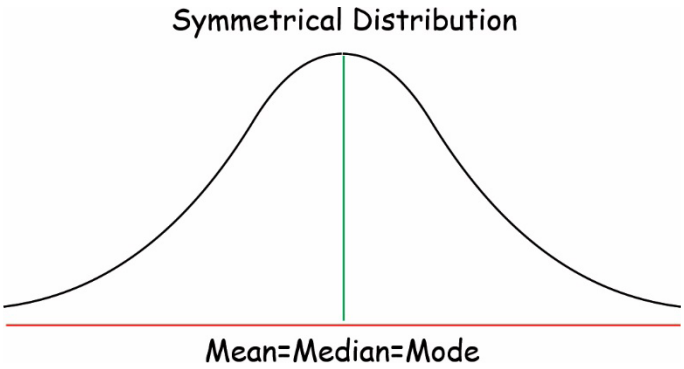


Figure 17: Symmetric data

3. **Asymmetric data:** A data which is not symmetric is known as asymmetric data. A measure of the degree of symmetry is called skewness. An asymmetric distribution will either be skewed to the right (positively skewed or right-tailed) or skewed to the left (negatively skewed or left-tailed)

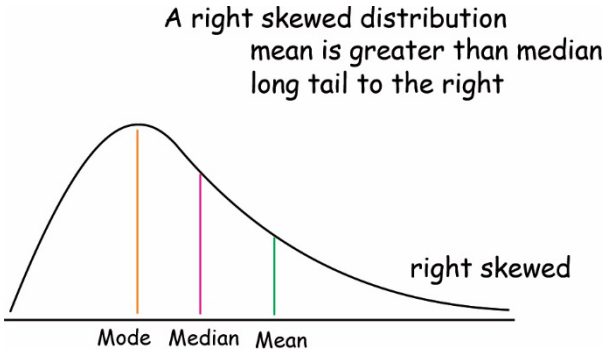


Figure 18: Right skewed

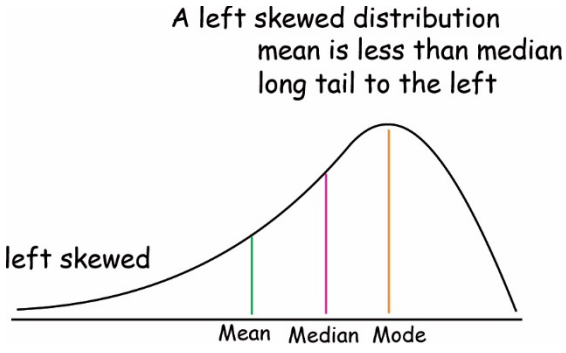


Figure 19: Left skewed

When is it appropriate to use the median as a measure of central tendency?

The mean is one of the most popular measures of central tendency but has one major flaw. Its value is influenced by outliers.

Example 30

Consider the weight of a sample of girls below:

60, 59, 63, 58, 60, 55, 56, 60, 62, 57.

Calculate the mean, median and mode.

Solution

$$\text{Mean} = \frac{60 + 59 + 63 + 58 + 60 + 55 + 56 + 60 + 62 + 57}{10} = \frac{590}{10} = 59$$

To find the median, we will first arrange the numbers in order.

55, 56, 57, 58, 59, 60, 60, 60, 62, 63

$$\text{Median} = \frac{59 + 60}{2} = \frac{119}{2} = 59.5$$

The mode from the distribution is 60

Therefore, mean ≈ median ≈ mode, so the data is fairly symmetrical.

How will the mean and median change if the entry of 62 was mistakenly recorded as 105?

The new data set will be 60, 59, 63, 58, 60, 55, 56, 60, 105, 57.

$$\text{Mean} = \frac{60 + 59 + 63 + 58 + 60 + 55 + 56 + 60 + 105 + 57}{10} = \frac{633}{10} = 63.3$$

This is significantly higher than 59.5, so the mean has been impacted by the outlier.

To find the median, we will first arrange the numbers in order.

55, 56, 57, 58, 59, 60, 60, 60, 63, 105

$$\text{Median} = \frac{59 + 60}{2} = \frac{119}{2} = 59.5$$

The median remains unchaged so it is not affected by the extreme value.

The mean is **sensitive** to outliers but the median is **resistant** to outliers.

Therefore, if the data is **NOT** symmetrical (either skewed to the left or right), the most appropriate measure of central tendency will be the median.

When to use the mean as a measure of central tendency

When the data is symmetrical, the mean and the median are close or almost the same. In this case it is appropriate to use the mean.

(Note, the data is said to be **perfectly symmetrical** when the mean is equal to the median.)

When to use the mode as a measure of central tendency

The mode is only used when the data is nominal or categorical. For instance, the data shows the preferred movie genres of Form 2 Home Economics Students of Ogyeedom SHTS.

Table 50: Frequency distribution

Movie Genre	Comedy	Action	Animation	Romance	Horror
Number of learners	15	3	10	20	12

In the above data, we cannot calculate the median since the order of the movie genre is not important. The data could have started with Horror, followed by Action etc. Therefore, mode is the only measure of central tendency that can be used in this instance. It would be appropriate to use if this meant choosing a genre of film to be enjoyed by the most people in the class.

Note, there are occasions when the median can be used for all categorical data. If the data is sequenced, we can calculate the median. For example, in days of the week, months of the year, educational levels, ranks in occupation, social status etc. Also, data with outcomes like *strongly agree, agree, disagree, and strongly disagree*. We can calculate the median for these types of data.

Example 31

The list below shows the number of times a sample of students were absent from school.

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

(a) Draw a frequency distribution table for the data.

(b) Calculate the (i) mean, (ii) median, (iii) mode

(c) What is the best measure of central tendency for the data? Justify your answer.

Solution

(a)

Table 51: Frequency distribution

Days Absent (x)	Number of learners(f)	*fx*	Cumulative frequency
0	6	0	6
1	7	7	13
2	8	16	21
3	6	18	27
4	5	20	32
5	1	5	33
Total	33	66	33

(b) i) mean $= \frac{66}{33} = 2$ days

ii) The number of observations is 33, the median will be the 17^{th} number.

From the table, the 17^{th} digit will be a 2

Therefore, the median = 2 days.

iii) Mode from the table = 2 days

(c) The most appropriate measure of central tendency for this data is the mean as this takes into account every data point and the data is symmetrical since the mean = median.

Example 32

The chart shows the occupations of a sample of residents of Kpasa District.

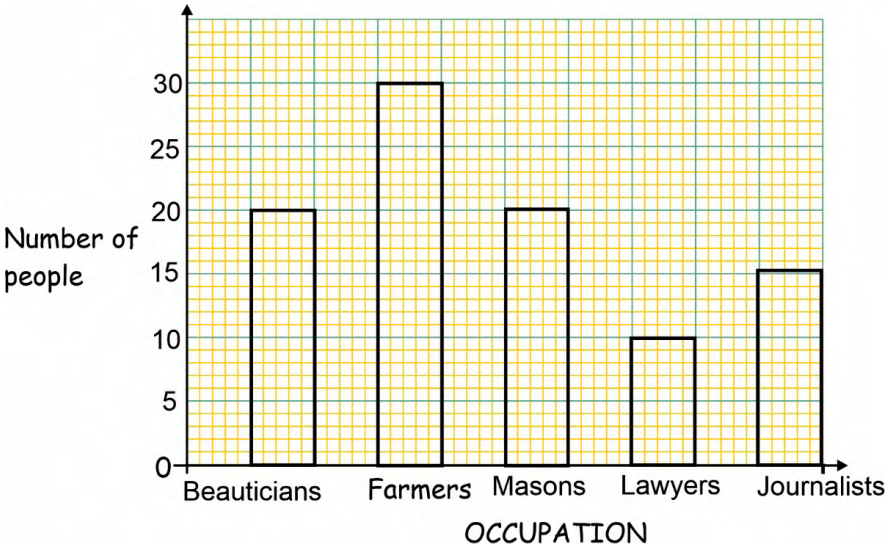


Figure 20: A bar graph showing occupations of some Kpasa Residents

Giving a reason, which method of central tendency will be appropriate for reporting this data?

Solution

The most appropriate measure of central tendency is the mode, because the data is categorical and cannot be sequenced.

Other Measures of Location

The mean, median and mode are examples of measures of location. Other measures of location are the **quartiles** and **percentiles**.

A **percentile** is a term that describes how a score compares with other scores from the same data set. The percentile is expressed as the percentage of values in a set of data that fall below a given value. If a value is in the r^{th} percentile, it implies the value is greater than r% of the total values in a data set. For example, if the 80^{th} percentile of ages of students in your school is 17 years, it means 80% of students in your school are aged below or equal to 17 years and 20% are aged above or equal to 17 years. Similarly, if your test score in an examination puts you at the 90^{th} percentile, it does not mean you scored 90% on the test, it means you performed better than 90% of students you took the same test with and only 10% outperformed you.

The **quartiles** are a subset of the percentiles. The quartiles are special percentiles. Quartiles divide a dataset into four equal parts, each containing 25% of the data.

The three quartile points are:

1. Lower quartile or first quartile (Q_1)

   The **lower quartile** or first quartile (Q1) is the median of the lower half of the dataset meaning it is the middle value of all data values whose positions are below the median. Like the percentile, if for instance, the lower quartile of a distribution is 10, it means 25% or a quarter, of the data values are less than or equal to 10.

2. Middle quartile or second quartile, also called the median. (Q_2)

   You can think of the median as the *middle value*, but it does not actually have to be one of the observed values. It is a number that separates ordered data into two halves. Half the values are the same number or smaller than the median and half the values are the same number or larger

3. Upper or third quartile (Q_3)

   It is the middle value of all values whose positions are above the median. It is also defined as the median of the upper half of the data. If the upper quartile of a given data is 10, it means 75% or three-quarters of the data values are less than or equal to 10. It also means 25% of the data values are greater than or equal to 10.

Comparing percentile to quartiles:

The 50^{th} percentile is equal to the median.

The 25^{th} percentile is the lower quartile.

The 75^{th} percentile is the upper quartile.

Example 33

This list gives the number of SHS in a sample of towns in Ghana.

19, 12, 20, 14, 16, 10, 19, 16, 15, 13 and 10

Calculate the median, lower quartile and upper quartile and interpret your answers.

Solution

We will first arrange the numbers in ascending order of magnitude.

10 10 12 13 14 15 16 16 19 19 20

There are a total of eleven numbers (eleven towns). The middle number is the 6^{th} value.

10	10	12	13	14	**15**	16	16	19	19	20
1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}	11^{th}

The median is 15.

Interpretation: The number of SHS in half of the sampled towns are less than or equal to 15 and the other half have SHS greater than or equal to 15.

For the lower quartile, we will find the middle value of all data values below the median position.

All data values below the median position are: 10, 10, 12, 13, 14

The median of the lower half of the data is 12. Therefore, the lower quartile is 12

Interpretation: The number of SHS in a quarter of the sampled towns is less than or equal to 12 and the number of SHS in the remaining three-quarters is greater than or equal to 12.

For the upper quartile, we will find the middle value of all data values above the median position.

All data values above the median position are: 16, 16, 19, 19, 20

The median of the upper half of the data is 19. Therefore, the upper quartile is 19.

Interpretation: The number of SHS in 75% of the sampled towns is less than or equal to 19 and the number of SHS in the remaining 25% of the towns is greater than or equal to 19.

Example 34

The following data shows the mathematics test results of 16 students.

61, 70, 75, 68, 80, 71, 58, 78, 85, 64, 83, 55, 82, 70, 67, 58

Calculate and interpret the

(i) median

(ii) lower quartile

(iii) upper quartile

Solution

(i) Arranging the data in ascending order:

55, 58, 58, 61, 64, 67, 68, 70, 70, 71, 75, 78, 80, 82, 83, 85

The data has 16 values. This means the median will be the average of the 8^{th} and 9^{th} values.

55	58	58	61	64	67	68	**70**	**70**	71	75	78	80	82	83	85
1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}	13^{th}	14^{th}	15^{th}	16^{th}

$$\text{Median} = \frac{70+70}{2} = \frac{140}{2} = 70$$

Interpretation: The mark of half of the students is less than or equal to 70 and the marks of the remaining half of the students is greater than or equal to 70.

(ii) In this case, the median position is neither the 8^{th} nor the 9^{th} position but the 8.5^{th} position. The Lower Quartile is the median of data value which falls *below* the median position.

Data values below the median position are

55	58	58	**61**	**64**	67	68	70

The median of these numbers is $= \frac{61+64}{2} = \frac{125}{2} = 62.5$.

Therefore, the lower quartile= 62.5.

Interpretation: The mark of 25% of the students is less than or equal to 62.5 and the marks of the remaining 75% of the students is greater than or equal to 62.5

(iii) The Upper Quartile is the median of the upper half of the data or all values above the median position. Data values above the median position are

70	71	75	78	**80**	82	83	85

The median of these numbers is $= \frac{78+80}{2} = \frac{158}{2} = 79$.

Therefore, the upper quartile = 79.

Interpretation: The mark of 75% of the students is less than or equal to 79 and the marks of the remaining 25% of the students is greater than or equal to 79

Example 35

The frequency table below gives the ages of students in a debate club.

Table 52: Frequency distribution

Age	11	12	13	14	15	16	17	18
Number of students	1	4	6	2	3	2	1	1

Calculate and interpret the (i) median (ii) lower quartile (iii) upper quartile

Solution

(i) Total frequency $= 1 + 4 + 6 + 2 + 3 + 2 + 1 + 1 = 20$

The median is the average of the numbers located at the 10^{th} and 11^{th} positions.

Table 53: Frequency distribution

Age	frequency	Position
11	1	1^{st} position is 11
12	4	2^{nd}, 3^{rd}, 4^{th} and 5^{th} positions are 12 each
13	6	6^{th} 7^{th} 8^{th} 9^{th} 10^{th} and 11^{th} positions are 13 each
14	2	12^{th} and 13^{th} positions are 14 each
15	3	14^{th} , 15^{th} and 16^{th} positions are 15 each
16	2	The 17^{th} and 18^{th} positions are 16 each
17	1	The 19^{th} position is 17
18	1	And the 20^{th} position is 18

From the table the 10^{th} and the 11^{th} data values are 13 and 13

We can also list the data values to have a clearer view:

11	12	12	12	12	13	13	13	13	**13**	**13**	14	14	15	15	15	16	16	17	18
1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}	13^{th}	14^{th}	15^{th}	16^{th}	17^{th}	18^{th}	19^{th}	20^{th}

$$\text{Median} = \frac{13 + 13}{2} = \frac{26}{2} = 13$$

Interpretation: Half of the members of the debate club are aged 13 or below and the remaining half are aged 13 or above.

(ii) The lower quartile is the data value at the median of the lower half of the data. Data values of the lower half of the data are the data occupying the 1^{st} to the 10^{th} positions. They are:

11, 12, 12, 12, **12, 13**, 13, 13, 13, 13

Lower Quartile$= \frac{12+13}{2} = \frac{25}{2} = 12.5$

Interpretation: 25% of members of the debate club are aged below or equal to 12.5 years and the remaining 75% are aged above or equal to 12.5 years.

(iii) The upper quartile is the data value at the median of the upper half of the data. Data values of the upper half of the data are the data occupying the 11^{th} to the 20^{th} positions. They are

13, 14, 14, 15, 15, 15, 16, 16, 17, 18

Upper Quartile$= \frac{15+15}{2} = \frac{30}{2} = 15$

Interpretation: 75% of members of the debate club are aged below or equal to 15 years and the remaining 25% are aged above or equal to 15 years.

Example 36

In a 100m race, the first quartile of the finishing time is 10seconds. Interpret the first quartile in the context of this situation.

Solution

25% percent of the competitors finished the race in 10 seconds or less and 75% of the competitors finished the race in 10 seconds or more.

Example 37

Aku's position in a dance competition placed her in the 85^{th} percentile. Interpret her position.

Solution

Aku performed better than 85% of her competitors and 15% of the competitors performed better than Aku.

Example 38

At a school, it was found that the upper quartile of the number of hours that students spend studying per week is 15. Interpret the upper quartile in the context of this situation.

Solution

75% of the students spend 15 hours or less studying per week, while 25% of the students spend 15 hours or more studying per week

MEASURES OF DISPERSION

When we disperse grains of rice or corn, we spread them. Likewise, measures of dispersion quantify the extent to which data values are spread. Measures of dispersion complement measures of central tendency by showing the degree of variability within a dataset. The measures of dispersion to be discussed include the range, standard deviation, variance and the interquartile range.

The interquartile range is used to compare the variability of two or more data sets, allowing the analyst to make relative assessments of the spread. Variance and standard deviation are widely used in finance to measure the unpredictability of asset returns by helping investors assess the risk associated with different investment options. In manufacturing and industrial processes, the range is used for quality control to monitor the variability of product specifications and ensure consistent quality by setting an acceptable range for certain dimensions or parameters.

Range

The range of a set of data is the difference between the largest and the smallest values in that dataset. The range gives us a rough idea of how widely the most extreme values in data are spread out. The range represents the interval that contains all the data values. A large range shows greater dispersion in the data while a small range shows less dispersion in the data. Because the range is calculated using only two data values, it is more useful with small data sets.

Range from an ungrouped data

Range = Largest value – Smallest value

Example 39

Find the range of the dataset, 19, 6, 7, 8, 20, 7, 10, 12 and 16.

Solution

The largest value is 20

The lowest value is 6

Range = 20 − 6 = 14.

Example 40

The prices (GH¢) of calculators in 12 shops at Makola market are:
180, 170, 169, 153, 174, 182, 166, 172, 155, 160, 173, 168

Calculate the range of prices.

Solution

The lowest or minimum price = GH¢153

The highest or maximum price = GH¢182

Range = 182 − 153 = 29

So, the range of prices of a calculator at Makola market is GH¢29

Range from grouped data

For grouped data, it is not possible to calculate the **exact** value of the range since we do not have the raw data. We however **estimate** the range using either of the methods below:

Range = (Upper Limit of the maximum class interval) – (Lower limit of the minimum class interval) = $\mathbf{U}_{max} - \mathbf{L}_{min}$

OR

Range = (Midpoint of the maximum interval) – (Midpoint of the minimum interval)

= $\mathbf{M}_{max} - \mathbf{M}_{min}$

Example 41

The distribution below represents goals scored by a sample of strikers in the Ghana Premier League in a soccer season.

Table 54: Frequency distribution

Goals	1 – 5	6 – 10	11 – 15	16 – 20	21 – 25	26 – 30
Number of players	4	6	8	6	4	2

Find the range.

Solution

We will use both methods to calculate the range.

Method 1

Range = (Upper Limit of the maximum class interval) – (Lower limit of the minimum class interval) = $\mathbf{U}_{max} - \mathbf{L}_{min}$

$\mathbf{U}_{max} = 30$

$\mathbf{L}_{min} = 1$

Range = $30 - 1 = 29$

Method 2

Range = (Midpoint of the maximum interval) – (Midpoint of the minimum interval)

$= \mathbf{M}_{max} - \mathbf{M}_{min}$

Midpoint of the maximum class interval = $\frac{30 + 26}{2} = \frac{56}{2} = 28$

Midpoint of the minimum class interval = $\frac{1 + 5}{2} = \frac{6}{2} = 3$

Range = $28 - 3 = 25$

Interquartile Range (IQR)

The inter-quartile range, also called **mid-spread,** measures the range of the middle 50% of a given data. Aside from being resistant to outliers, IQR is used in business as a marker for income rates. It is also used when building box plots.

IQR = Upper Quartile – Lower Quartile = $Q_3 - Q_1$

Note: The interquartile range is calculated with results obtained from lower and upper quartiles. In themselves, the quartiles and percentiles are not measures of dispersion but measures of location. However, as soon as we look at ranges we are looking at measures of dispersion.

Example 42

This list gives the number of surgeons in 19 major hospitals in Ghana:
23, 13, 20, 6, 11, 13, 25, 8, 10, 13, 12, 6, 23, 9, 12, 7, 13, 14 and 10
Calculate the inter-quartile range.

Solution

Arrange the numbers in ascending order.
6, 6, 7, 8, 9, 10, 10, 11, 12, 12, 13, 13, 13, 13, 14, 20, 23, 23, 25
Find the median as this will gauge where to locate the lower and the upper quartiles.
The median is the middle number.
6, 6, 7, 8, 9, 10, 10, 11, 12, **12**, 13, 13, 13, 13, 14, 20, 23, 23, 25
Median is 12

The **Lower Quartile** is the median of all numbers which are positioned *below* the median. From the ordered list, numbers positioned below the median are
6, 6, 7, 8, **9**, 10, 10, 11, 12
The median of these numbers is 9. Therefore, the Lower Quartile = 9

The **Upper Quartile** is the median of all numbers which are positioned *above* the median. From the ordered list, numbers positioned above the median are
13, 13, 13, 13, **14**, 20, 23, 23, 25
The median of these numbers is 14. Therefore, the Upper Quartile = 14
The interquartile range = Upper Quartile – Lower Quartile = 14 – 9 = 5

Example 43

The following data shows the mathematics test results of 20 students.
25, 39, 28, 30, 20, 35, 20, 24, 27, 19, 24, 36, 17, 36, 38, 18, 38, 26, 23, 17
Calculate the interquartile range.

Solution

Arrange the data in ascending order:

17, 17, 18, 19, 19, 20, 20, 23, 24, 24, 25, 26, 27, 28, 30, 35, 36, 36, 38, 38, 39

The data has 20 values. This means the median will be the average of the 10^{th} and 11^{th} values.

17	17	18	19	19	20	20	23	24	24	25	26	27	28	30	35	36	38	38	39
1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}	13^{th}	14^{th}	15^{th}	16^{th}	17^{th}	18^{th}	19^{th}	20^{th}

$$\text{Median} = \frac{24 + 25}{2} = \frac{49}{2} = 24.5$$

In this example, the median position is neither the 10^{th} nor the 11^{th} position but rather the 10.5^{th} position.

The lower quartile is the median value of the lower half of the data

Data values below the median position are

17	17	18	19	19	20	20	23	24	24

The median of these numbers is $= \frac{19 + 20}{2} = \frac{39}{2} = 19.5$

Therefore, the lower quartile= 19.5

The Upper Quartile is the median value of the upper half of the data

Data values above the median position are

25	26	27	28	30	35	36	38	38	39

The median of these numbers is $= \frac{30 + 35}{2} = \frac{65}{2} = 32.5$

Therefore, the upper quartile $= 32.5$

Interquartile range $= 32.5 - 19.5 = 13$

Example 44

The data set shows the monthly salary ($) of 13 managers.

4100, 4100, 8000, 1800, 4300, 2000, 4700, 2500, 3800, 2000, 4700, 4500 and 3200

(a) Why should the interquartile range be used to describe the spread of this data?

(b) Calculate the interquartile range.

(c) Each manager received a 5% salary cut. How will this affect the interquartile range (IQR)?

Solution

(a) The interquartile range should be used because the data contains extreme values (outliers)

(b) Arrange the data in ascending order:

1800, 2000, 2000, 2500, 3200, 3800, **4100**, 4100, 4300, 4500, 4700, 4700, 8000

The median is 4100

The lower quartile is the median of 1800, 2000, **2000**, **2500**, 3200, 3800

$$\text{Lower Quartile} = \frac{2000 + 2500}{2} = \frac{4500}{2} = 2250$$

The upper quartile is the median of 4100, 4300, **4500**, **4700**, 4700, 8000

$$\text{Upper Quartile} = \frac{4500 + 4700}{2} = \frac{9200}{2} = 4600$$

Therefore, interquartile range $= 4600 - 2250 = 2350$

(c) If each manager receives a salary cut, the IQR will get smaller. When each manager takes a 5% salary cut, it means their new salary will be 95% of their old salary, so the IQR will also be cut by 5%. To demonstrate this:

$$\begin{aligned}\text{New interquartile range} &= \frac{95}{100} \times 4600 - \frac{95}{100} \times 2250 \\ &= 4370 - 2137.5 \\ &= 2232.5\end{aligned}$$

This is 95% of the old IQR ($0.95 \times 2350 = 2232.5$)

Example 45

Construct a cumulative frequency curve for the data below and estimate the quartiles from the cumulative frequency curve and calculate the IQR.

Table 55: Frequency distribution

Weight (kg)	Frequency
$50 \leq w < 55$	4
$55 \leq w < 60$	13
$60 \leq w < 65$	21
$65 \leq w < 70$	34
$70 \leq w < 75$	48
$75 \leq w < 80$	27
$80 \leq w < 85$	18
$85 \leq w < 90$	9
$90 \leq w < 95$	5
$95 \leq w < 100$	2

Solution

Table 56: Frequency distribution

Weight (kg)	Frequency	Upper boundary	Cumulative frequency
$50 \leq w < 55$	4	55	4
$55 \leq w < 60$	13	60	17
$60 \leq w < 65$	21	65	38
$65 \leq w < 70$	34	70	72
$70 \leq w < 75$	48	75	120
$75 \leq w < 80$	27	80	147
$80 \leq w < 85$	18	85	165
$85 \leq w < 90$	9	90	174
$90 \leq w < 95$	5	95	179
$95 \leq w < 100$	2	100	181

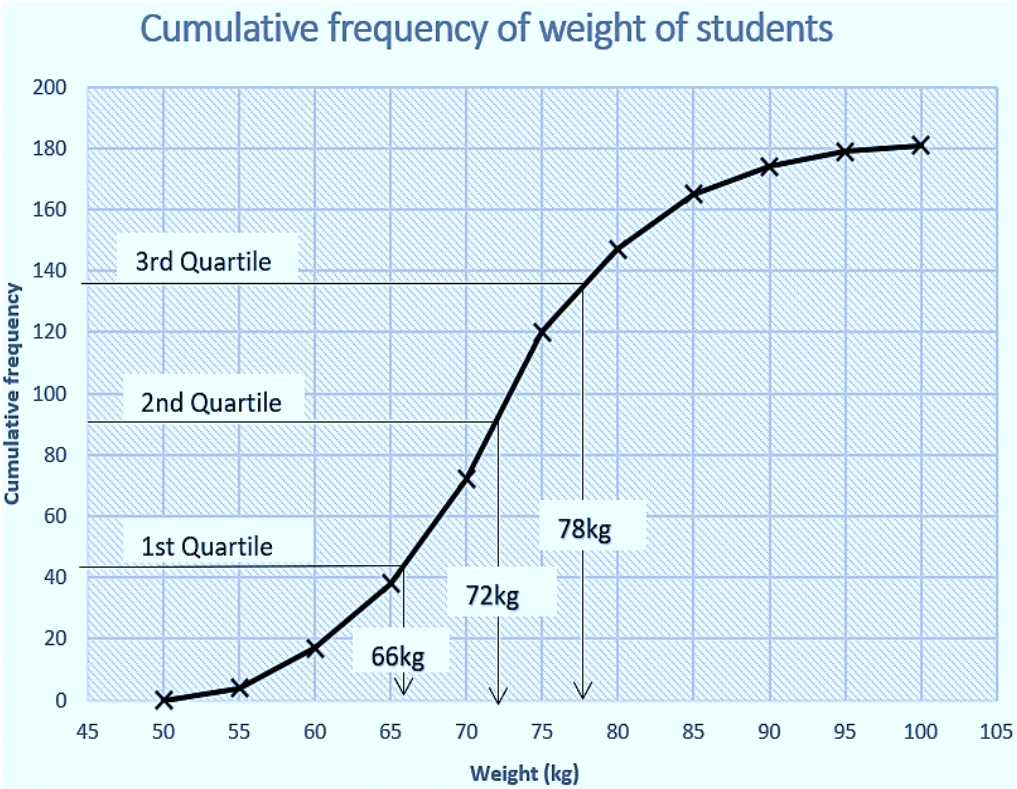


Figure 21: A cumulative frequency curve showing the weight of students

From the curve:

Cumulative frequency $N = 181$

Thus, $1st\ Quartile\ (Q_1) = \frac{181+1}{4} = 45.5^{th}\ observation,\ Q_1 = 66kg$

$2nd\ Quartile\ (Q_2) = \frac{181+1}{2} = 91^{st}\ observation,\ Q_2 = 72kg$

$3rd\ Quartile\ (Q_3) = \frac{3(181+1)}{4} = 136.5^{th}\ observation,\ Q_3 = 78kg$

Note: The frequency (cumulative frequency, CF) is the location where the quartiles are traced from and are not the actual quartiles. The CF is the number of persons thus we divide the frequency by 4 and not the weight.

Standard Deviation and Variance

The **standard deviation** is a statistic that tells us the spread of data about the mean. Standard deviation tells the variation in the data. It tells us how measurements for a group are spread out from the mean. A low standard deviation indicates that the data values tend to be close to the mean. A high standard deviation shows that the data values are spread out over a wider range of values. The standard deviation is denoted by the Greek letter σ (*sigma*) or the Latin letter s.

Variance is the square of the standard deviation. It is often represented as σ^2 or s^2 or $Var(X)$.

Formula for calculating standard deviation

If $x_1, x_2, \ldots\ldots, x_n$ are n individual numbers (ungrouped data), then

$$SD = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} \quad or \; SD = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$$

If x_1 has frequency f_1,

x_2 has frequency f_2,

x_3 has frequency f_3 and so on up to x_n with frequency f_n, then,

$$SD = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}} \; or \; SD = \sqrt{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2}$$

Where x is class marks or mid-point in cases where the data is grouped.

n =total number of data values.

$\sum f$ = total frequency

Calculating Standard Deviation using Assumed Mean Method

In situations where we have ungrouped data. Note that $d = x - A$ with A being the assumed mean and n = total number of numbers.

$$SD = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

For grouped data:

$$SD = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum d}{\sum f}\right)^2}$$

Example 46

The age of a sample of retirees are 90, 61, 80, 60, 55, 70, 68 and 76.
Calculate the standard deviation.

Solution

Using $SD = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$

We have mean $(\bar{x}) = \frac{90+61+80+55+70+68+76}{8} = \frac{560}{8} = 70$

$$\sum(x-\bar{x})^2 = \left\{ \begin{matrix} (90-70)^2 + (61-70)^2 + (80-70)^2 + (60-70)^2 \\ +(55-70)^2 + (70-70)^2 + (68-70)^2 + (76-70)^2 \end{matrix} \right\}$$

$$= \{(20)^2 + (-9)^2 + (10)^2 + (-10)^2 + (-15)^2 + (0)^2 + (-2)^2 + (6)^2\}$$
$$= 400 + 81 + 100 + 100 + 225 + 0 + 4 + 36 = 946$$

$$SD = \sqrt{\frac{946}{8}} = \sqrt{118.25} = 10.87$$

Alternatively,

Using $SD = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$ will also give the same result.

$$\sum x^2 = 90^2 + 61^2 + 80^2 + 60^2 + 55^2 + 70^2 + 68^2 + 76^2 = 40146$$

$$\sum x = 90 + 61 + 80 + 60 + 55 + 70 + 68 + 76 = 560$$

$$SD = \sqrt{\frac{40146}{8} - \left(\frac{560}{8}\right)^2}$$

$$SD = \sqrt{5018.25 - 4900}$$

$$SD = \sqrt{118.25} = 10.87$$

Example 47

The data below is the masses of ten water tanks:
6.7, 10, 9.3, 5.9, 9.4, 6.8, 7.6, 7.8, 8.0 and 8.5
Using assumed mean of 7, calculate

(a) Standard deviation

(b) variance

Solution

(a) Formula for standard deviation is :$SD = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$

The deviations (d) from the mean are:

$x - A$	=	d
$6.7 - 7$	=	-0.3
$10 - 7$	=	3
$9.3 - 7$	=	2.3
$5.9 - 7$	=	-1.1
$9.4 - 7$	=	2.4
$6.8 - 7$	=	-0.2
$7.6 - 7$	=	0.6
$7.8 - 7$	=	0.8
$8.0 - 7$	=	1
$8.5 - 7$	=	1.5

$\sum d = -0.3 + 3 + 2.3 + (-1.1) + 2.4 + (-0.2) + 0.6 + 0.8 + 1 + 1.5 = 10$

$\sum d^2 = (-0.3)^2 + 3^2 + 2.3^2 + (-1.1)^2 + 2.4^2 + (-0.2)^2 + 0.6^2 + 0.8^2 + 1^2 + 1.5^2$

$= 25.64$

$SD = \sqrt{\frac{25.64}{10} - \left(\frac{10}{10}\right)^2} = \sqrt{2.564 - 1} = \sqrt{1.564} = 1.25$

(b) Variance $= (SD)^2 = \left(\sqrt{1.564}\right)^2 = 1.564$

Example 48

The deviations of the ages of students from 15 are
$-6, -5, -4, 6, 8, 7, -1, 2, 5, -3, 3$ and 12.

Calculate:

(a) Standard Deviation

(b) Variance

Solution

(a)

$$\sum d = -6 - 5 - 4 + 6 + 8 + 7 - 1 + 2 + 5 - 3 + 12 = 24$$

$$\sum d^2 = \{(-6)^2 + (-5)^2 + (-4)^2 + 6^2 + 8^2 + 7^2 + (-1)^2 + 2^2 + 5^2 + (-3)^2 + 3^2 + 12^2\}$$

$$= 418$$

$$SD = \sqrt{\frac{418}{12} - \left(\frac{24}{12}\right)^2} = \sqrt{34.8333 - 4} = \sqrt{30.8333} = 5.5528$$

(b) Variance $= (S.D)^2 = 5.5528^2 = 30.83$

Example 49

The table below shows the distribution of heights of students in a school

Table 57: Frequency distribution

Heights(m)	1.2 – 1.26	1.27 – 1.33	1.34 – 1.40	1.41 – 1.47	1.48 – 1.54	1.55 – 1.61
Number of students	4	8	12	16	12	8

Calculate the standard deviation and the variance.

Solution

Table 58: Frequency distribution

Heights (m)	*f*	*x*	*fx*	fx^2
1.20 – 1.26	4	1.23	4.92	6.0516
1.27 – 1.33	8	1.30	10.40	13.52
1.34 – 1.40	12	1.37	16.44	22.5228
1.41 – 1.47	16	1.44	23.04	33.1776
1.48 – 1.54	12	1.51	18.12	27.3612
1.55 – 1.61	8	1.58	12.64	19.9712
	60		85.56	122.6044

$$SD = \sqrt{\frac{122.6044}{60} - \left(\frac{85.56}{60}\right)^2} = \sqrt{2.04341 - 2.033476}$$

$$= \sqrt{0.00993} = 0.09965 = 0.10$$

Variance $= 0.00993$

Example 50

Use assumed mean of 1.44 to solve example 49.

Solution

Table 59: Frequency distribution

Heights (m)	*f*	*x*	$d = x - 1.44$	*fd*	fd^2
1.20 – 1.26	4	1.23	−0.21	−0.84	0.1764
1.27 – 1.33	8	1.30	−0.14	−1.12	1.1568
1.34 – 1.40	12	1.37	−0.07	−0.84	0.0588
1.41 – 1.47	16	1.44	0	0	0
1.48 – 1.54	12	1.51	0.07	0.84	0.0588
1.55 – 1.61	8	1.58	0.14	0.12	0.1568
	60			−0.84	0.6076

$$SD = \sqrt{\frac{0.6076}{60} - \left(\frac{-0.84}{60}\right)^2}$$

$$= \sqrt{0.0101266667 - 0.000196}$$

$$= \sqrt{0.00993}$$

$$= 0.09965 \approx 0.10$$

Variance = 0.00993

REVIEW QUESTIONS

Review Questions 9.1

1. Outline three reasons why statistics is important in real-life.
2. Differentiate between a questionnaire and an interview as methods of collecting data.
3. The headmistress of Mfantsiman Girls Senior High School wants to evaluate the overall performance of students in the school from Year 1 to Year 3. What sampling technique should she employ to ensure that every year group and class is fairly represented in the study?
4. Determine whether or not data from the following will be continuous or discrete:
   a. The number of broken chairs in different classrooms in a school.
   b. The monthly income of workers in a factory.
   c. Marks obtained by a student in a science examination.
   d. The distance travelled by a non-resident teacher to school each school day.
   e. The number of females in a technical class.
   f. Ages (years) of non-teaching staff in a school.
5. The weights (in kilograms) of 25 students in a classroom were measured and recorded as follows: 43 46 45 50 44 52 56 65 64 58 60 65 63 62 44 50 58 44 56 58 64 46 52 47 42

   Construct a frequency table for the data, showing the weight range and its corresponding frequency count.
6. Consider the following data collected from a survey about sales per day (GH¢) made by food vendors on campus: 300, 350, 250, 480, 320, 300, 250, 480, 320, 450, 450.
   a. Determine which scale of measurement (nominal, ordinal, interval, ratio) is most appropriate for categorising the sales of vendors on campus.

b. Justify your choice of scale based on the data's characteristics and the chosen scale.

7. The data below shows the number of hours allotted for personal study each week by Kojokrom Senior High School students. Put the data into a grouped frequency distribution table using intervals of 3 hours. Include on the table the class midpoint, class boundaries, class width and the corresponding frequency count.

6, 20, 13, 4, 20, 17, 13, 12, 8, 20, 7, 5, 12, 9, 12, 14, 4, 11, 10, 9, 5, 15, 7, 3, 17, 13, 17, 7, 8, 11, 18, 9, 19, 8, 11, 18, 10, 6, 19, 7, 19, 16, 18, 13, 4, 17, 19, 15, 13, 6, 6, 3, 19, 17, 11, 18, 14, 20, 9, 7.

8. The data below shows the marks of students on a test.

10	14	12	13	15	12	11	12	13	11
15	12	10	15	11	13	12	13	10	14
11	12	14	13	14	10	15	12	13	15
15	11	12	10	15	11	12	10	15	11

a. Draw a frequency distribution table for data.

b. Draw a bar chart for the data.

9. The list below shows the price (GH¢) of mathematics textbooks in a sample of shops.

133	132	134	133	132	133	134	133
132	133	131	131	133	135	133	132
133	135	133	132	133	131	132	133

a. Draw a frequency distribution table for the data.

b. Draw a pie chart to represent the data.

c. Find the fraction of shops which priced mathematics textbooks at GH¢133

10. The table below shows the daily expenditure for a family for weekdays.

Expenditure	Monday	Tuesday	Wednesday	Thursday	Friday
Amount (GH¢)	400	250	400	250	500

a. Draw a pie chart for the data.

b. What is the family's average daily expenditure.

11. The pie chart below shows the time learners reported into school.

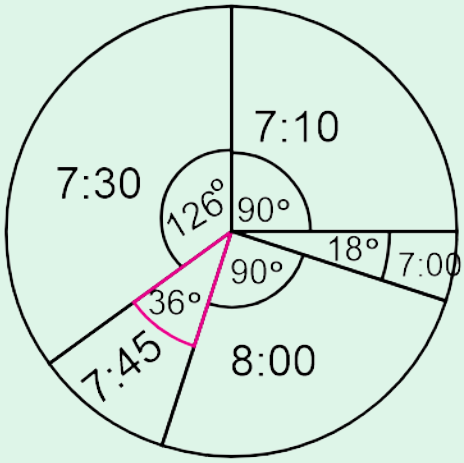


Figure 22: Pie chart showing reporting time of learners

a. Use the chart to complete the table below.

Reporting time (am)	Number of learners	Angle of sector
7:00	6	
7:10		
7:30		
7:45		
8:00		

b. What percentage of the learners reported at 7:30 am?

c. How many learners were late if reporting time to school was 7:30?

12. The table below shows the height of trees in a forest.

Height (m)	5	6	7	8	9	12
Number of trees	4	5	3	10	5	3

Calculate the mean height of the trees.

13. Calculate the mean of all prime numbers between 50 and 100.

14. Calculate the mean of the following set of numbers

a. 15, 18, 13, 6, 0, 1, 8, 15, 11, 9, 1, 3 and 13

b. 80, 90, -50, 100, 70, 48, 62, 23, and -20

15. The table below shows Kofi's sales for the week

Day	Monday	Tuesday	Wednesday	Thursday	Friday
Sales (GhC)	300	280	400	500	520

a. Calculate his total sales in the week

b. Calculate his mean daily sales.

16. The data below gives the number of hours a sample of teachers spends marking over the weekend.

5	3	5	5	3	4	4	6	5	7
3	4	4	6	4	5	3	4	3	4
4	6	7	3	6	3	5	7	4	3

a. Make a frequency distribution table for the data

b. Use your distribution table to calculate the mean.

17. In each of the sets of data, state the outliers and describe the shape of the distribution.

(a) 76, 66, 83, 75, 76, 5, 85, 72, 9, 76

(b) 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190

(c) 10, 1, 0, 1, 1, 0, 1, 2, 15, 2

(d) 17, 2, 16, 17, 3, 15, 16, 15, 3, 16

18. The data shows how a sample of teachers commute to work.

Mode of transportation	Cycle	Trotro	Taxi	Walking	School bus	Private car
Number of teachers	8	15	4	22	10	2

Which measure of central tendency will be most appropriate for the data? Justify your response.

19. The data shows the number of libraries in a sample of twenty towns

3	2	2	3	1	7	2	3	1	2
2	6	1	2	3	2	1	2	7	1

(a) Draw a frequency distribution table for the data

(b) State the best method of central tendency to report findings on the data. Justify your answer.

20. The list shows the ages of a sample of animals.

23	21	23	21	23	21	25	19	27	23
25	19	25	23	5	23	9	23	23	25
7	23	21	21	25	25	23	19	25	25

(a) Draw a frequency distribution table for the data and use it to draw a bar chart

(b) Find the mode

(c) Calculate the mean and the median

(d) Which method of central tendency will you use to report findings on the data?

Justify your response.

21. The data below shows the time, t minutes, taken by a sample of students to complete a mathematics homework.

36	60	43	58	33	48	43	54	34	64
45	52	48	60	40	54	64	49	43	56
56	61	34	31	52	43	50	40	53	40
32	58	42	64	53	38	41	64	43	45
45	53	52	48	65	35	31	33	32	55
65	39	37	35	43	49	40	43	56	42

a. Prepare a frequency distribution table using 31-35, 36-40, 41-45, etc

b. Draw a histogram to represent the data.

22. The list below gives the career goals of a sample of female footballers.

254	263	255	239	259	252	259	265
251	245	249	238	246	256	251	254
241	247	240	256	261	251	264	253
255	253	257	260	248	258	236	244
242	248	258	250	252	249	252	257
265	237	252	260	243			

(a) Copy and complete the frequency distribution table below

Carrier goals	Tally	Number of students
236 – 240		
241 – 250		
251 – 253		
254 – 260		
261 – 265		

(b) Draw a histogram for the data.

23. The data blow shows the pocket money of a sample of day students at Achimota School

Amount	Number of students
$10 < GH¢ \leq 24$	21
$24 < GH¢ \leq 40$	16
$40 < GH¢ \leq 60$	40
$60 < GH¢ \leq 70$	25
$70 < GH¢ \leq 100$	15
Total	

Draw a histogram to represent the data.

24. A policewoman records the speed of a sample of 132 vehicles on a road with a speed limit of 70km per hour. The histogram below represents the results.

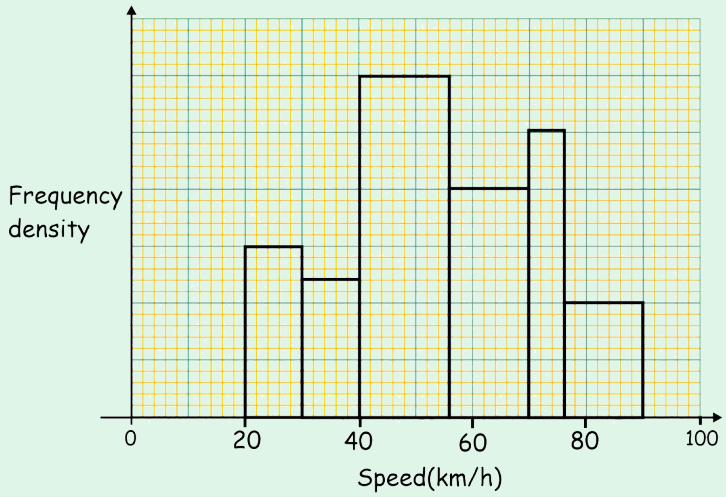


Figure 23: A histogram based on speed records of a policewoman

(a) Construct a frequency distribution table for the data.

(b) Calculate the number of vehicles that were travelling below the speed limit by at least 14km/h

(c) Reconstruct the histogram to include the frequency density.

(d) If each driver who exceeded the speed limit was fined GH¢ 400.00, find the total amount accrued from fines.

25. The histogram below shows information about the age of children who participated in a research study about mental health.

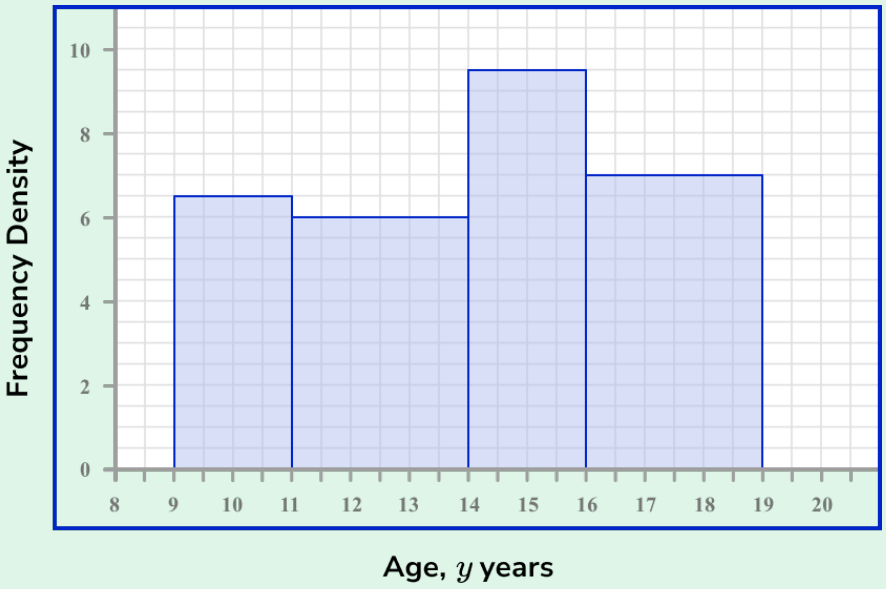


Figure *24*: A histogram showing age of children in a research study

(a) Use the data draw a frequency distribution table.

(b) Estimate the number of students between 12 and 14 years old who participated in the research study.

26. The data below shows the life expectancy (in years) of a sample of animals.

11	3	22	18	14	15	17	16	29	6
24	27	12	23	28	5	22	20	18	16
2	24	28	23	21	15	27	14	31	32
19	18	32	3	17	14	9	11	26	21
21	27	17	34	24	32	17	32	21	18
27	22	28	38	14	18	29	19	8	19
11	37	7	22	33	37	33	20	26	11
14	16	39	11	23	5	19	14	22	4

(a) Using class intervals of 1 – 5, 6 – 10, 11 – 15, etc. Construct a cumulative frequency distribution table for the data.

(b) Draw a cumulative frequency curve for the distribution.

27. The table below shows the time (hours) it took a sample of participants to complete a challenge.

Time (hours)	2.5 – 3	3.1 – 3.5	3.6 – 4	4.1 – 4.5	4.6 – 5	5.1 – 5.5	5.6 – 6
Number of participants	3	4	6	10	18	9	4

(a) Draw a cumulative frequency curve for the data

(b) Calculate the mean mark

28. The cumulative frequency curve gives information about the height (m) of a sample of trees in Dodowa forest.

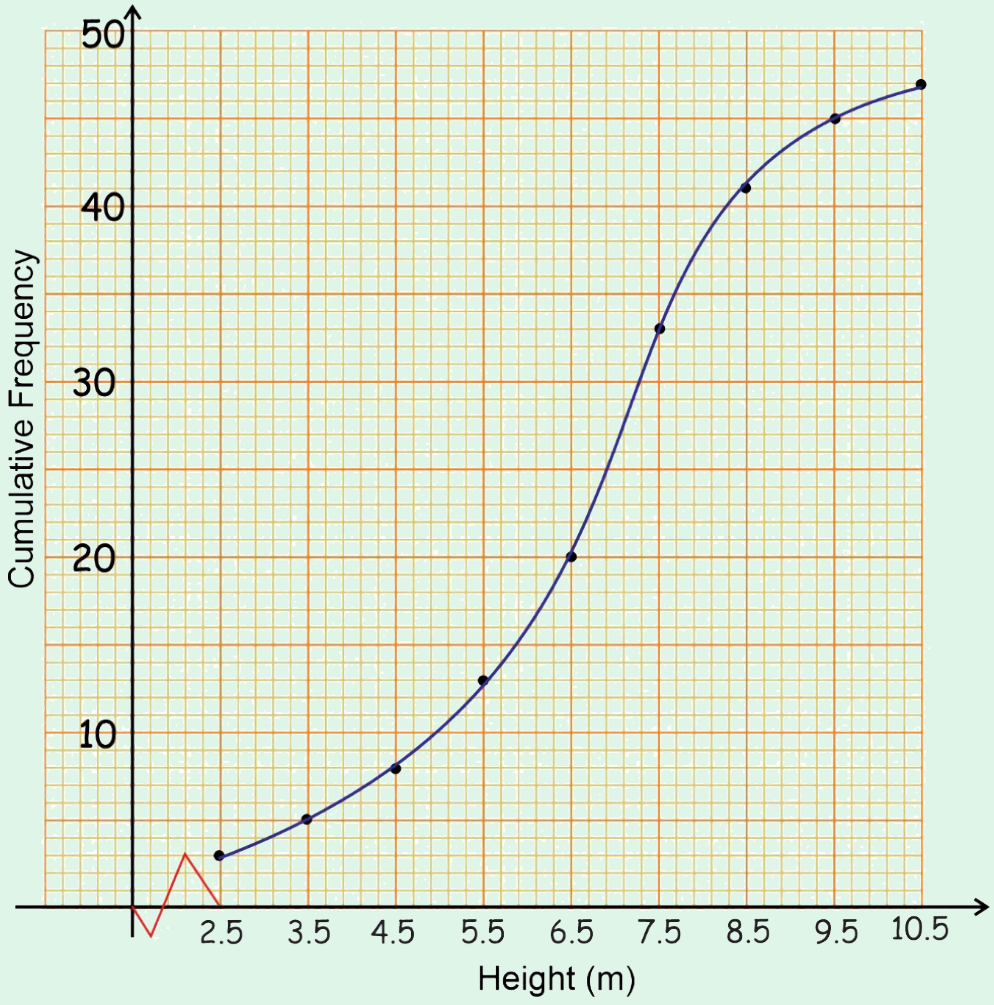


Figure 25: Cumulative frequency on height of trees in Dodowa forest.

(a) Use the graph to copy and complete the table below

Class boundaries	Cumulative frequency	Frequency
1.5 – 2.5	3	3
2.5 – 3.5	5	2
3.5 – 4.5	8	3
4.5 – 5.5	13	5
5.5 – 6.5	20	7
6.5 – 7.5	33	13
7.5 – 8.5	41	8
8.5 – 9.5	45	4
9.5 – 10.5	47	2

(b) Use the distribution table to calculate the mode, correct to 2 significant figures.

(c) Calculate the mean age.

29. The following data represent the number of employees at various hotel in Accra.

22, 35, 15, 26, 40, 28, 18, 20, 25, 34, 39, 42, 24, 22, 19, 27, 22, 34, 40, 20, 38, 28, 23

Use the data to find (a) the median (b) Lower Quartile (c) Interquartile range

30. For the following 13 bicycle prices (GH¢), calculate the Interquartile Range and the standard deviation

389, 230, 158, 479, 639, 114, 550, 387, 659, 529, 575, 488, 700

31. Fifty mothers were asked how much sleep they get per day (rounded to the nearest hour). The results were as follows:

Amount of Sleep per night	Frequency
6	2
7	5
8	7
9	12
10	14
11	10

Find

(i) the interquartile range

(ii) the variance and the standard deviation

ANSWERS TO REVIEW QUESTIONS

1. Many reasons, but some include:
   a. To help make informed decisions
   a. For fact checking what you are told
   b. To help validate scientific research hypotheses
2. A questionnaire is a written set of questions that respondents answer on their own, often distributed through paper, email or online forms. An interview is a verbal exchange where the interviewer asks questions and records the responses, either in person, over the phone or through a video call.
3. Stratified sampling technique.
4. a. Discrete
   b. Continuous
   c. Discrete
   c. Continuous
   d. Discrete
   e. Discrete
5.

Weight (kg)	Frequency (f)
41 – 45	6
46 – 50	5
51 – 55	2
56 – 60	6
61 – 65	6
Total	**25**

6. a. Ratio
   b. To be able to make sales a commodity has to be sold and paid for as such recording zero means no sales have been made thus zero is

absolute for sales. Since ratio scale is the only scale with an absolute/true zero, sales are on a ratio scale.

7.

Personal study time (hours)	Class midpoint	Frequency (f)	Class boundaries	Class width
3 – 5	4	7	$2.5 \leq x < 5.5$	3
6 – 8	7	12	$5.5 \leq x < 8.5$	3
9 – 11	10	10	$8.5 \leq x < 11.5$	3
12 – 14	13	10	$11.5 \leq x < 14.5$	3
15 – 17	16	8	$14.5 \leq x < 17.5$	3
18 – 20	19	13	$17.5 \leq x < 20.5$	3
Total		**60**		

8. (a)

Marks	Tally	Frequency
10	𝍸/	6
11	𝍸//	7
12	𝍸////	9
13	𝍸/	6
14	////	4
15	𝍸////	8
Total		**40**

(b)

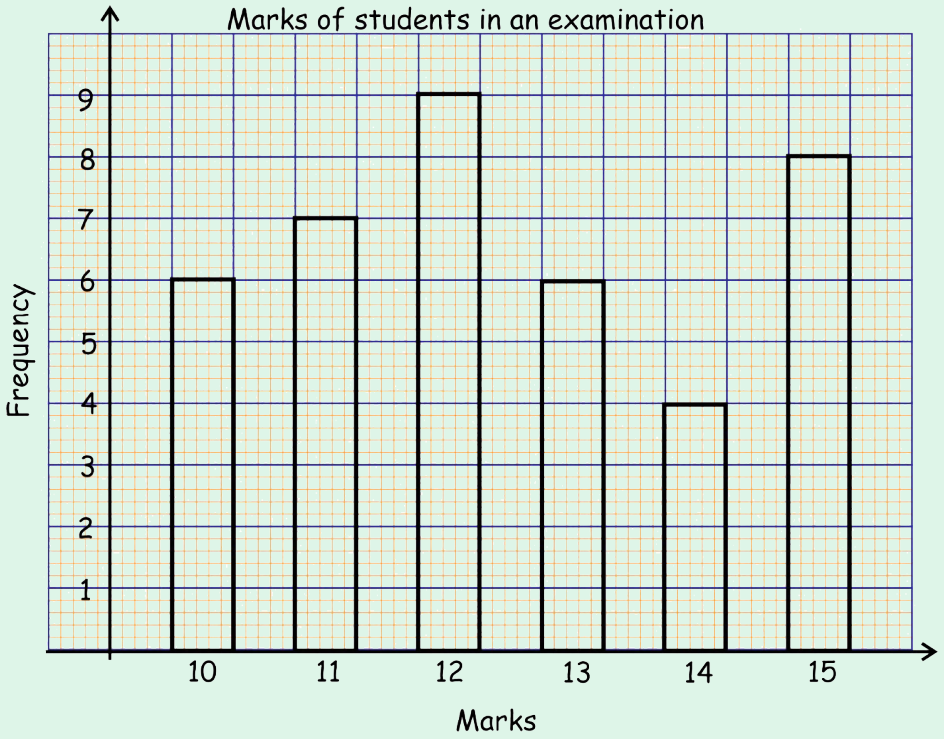


Figure 26: A bar graph showing marks in an examination.

9.

a)

Price/GH¢	Frequency
131	3
132	6
133	11
134	2
135	2
Total	**24**

b)

Price	Angle of Sector
131	$\frac{3}{24} \times 360° = 45°$
132	$\frac{6}{24} \times 360° = 90°$
133	$\frac{11}{24} \times 360° = 165°$
134	$\frac{2}{24} \times 360° = 30°$
135	$\frac{2}{24} \times 360° = 30°$

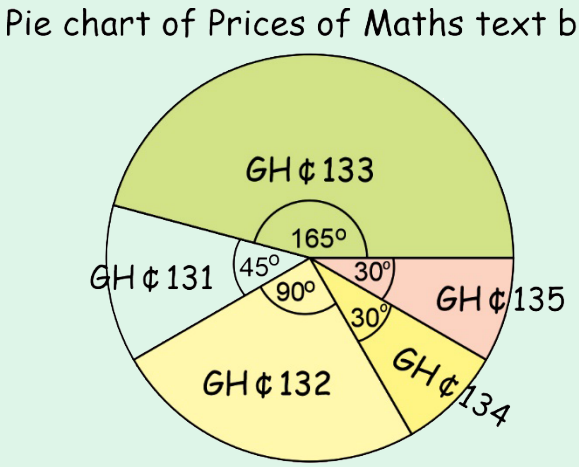


Figure 27: A pie chart showing prices of textbooks.

c) Fraction of shops which priced mathematics textbook at GH¢ 133 $= \frac{11}{24}$

10.

a)

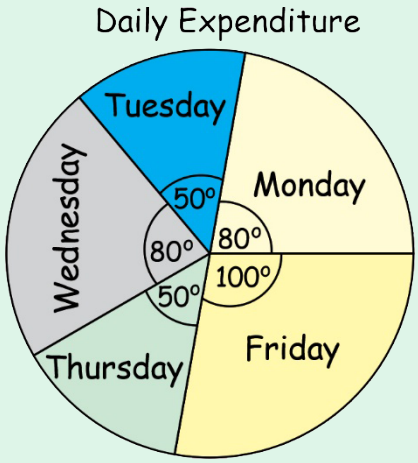


Figure 28: A pie chart showing daily expenditure

b) Average daily expenditure $= \frac{400 + 250 + 400 + 250 + 500}{5} = €\ 360$

11. (a)

Reporting time (am)	Number of learners	Angle of sector
7:00	6	18°
7:10	30	90°
7:30	42	126°
7:45	12	36°
8:00	30	90°

b) 35% of learners reported at 7.30am.

c) 42 learners were late.

12. Mean = 7.73m

13. mean $= \frac{53 + 59 + 61 + 67 + 71 + 73 + 79 + 83 + 89 + 97}{10} = \frac{732}{10} = 73.2$

14. **(a)** 8.69

(b) 44.78

15. **(a)** Total Sales = GH¢2000

(b) Mean Sales = GH¢400

16. **(a)**

Time (hours)/x	Tally	Number of teachers/f	fx
3	## ///	8	24
4	## ////	9	36
5	## /	6	30
6	////	4	24
7	///	3	21
Total		30	135

(b) Mean = 4.5 hours = 4 hours 30 minutes.

17. **(a)** 76, 66, 83, 75, 76, 5, 85, 72, 9, 76

Outliers are 9 and 5. The distribution is skewed to the left or negatively skewed.

(b) 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190.

There are no outliers, the distribution is symmetrical.

(c) 10, 1, 0, 1, 1, 0, 1, 2, 15, 2

The outliers are 15 and 10. The distribution is skewed to the right or positively skewed.

(d) 17, 2, 16, 17, 3, 15, 16, 15, 3, 16

The outliers are 2 and 3. The distribution is negatively skewed.

18. The mode is the most appropriate as the data is categorical and not sequenced.

19. (a)

Number of libraries	1	2	3	4	5	6	7
Number of towns	5	8	4	0	0	1	2

(b) The median is the most appropriate as the distribution is positively skewed or skewed to the right since the mean is greater than the median.

20. (a)

Age	5	7	9	19	21	23	25	27
Frequency	1	1	1	3	5	10	8	1

(b) Mode = 23

(c) Mean = 21.333 years = 21years 4 months; Median = 23

(d) Most appropriate measure of location is the median as the data is negatively skewed.

21. (a)

Time(minutes)	Number of students
31 – 35	10
36 – 40	8
41 – 45	13
46 – 50	6
51 – 55	8
56 – 60	8
61 – 65	7

(b)

Figure 29: A histogram showing time spent

22. (a)

Carrier goals	Tally	Number of students	Class interval	Class width	Frequency density
236 – 240	𝍸	5	$235.5 \leq x < 240.5$	5	1
241 – 250	𝍸 𝍸 //	12	$240.5 \leq x < 250.5$	10	1.2
251 – 253	𝍸 ////	9	$250.5 \leq x < 253.5$	3	3
254 – 260	𝍸 𝍸 ////	14	$253.5 \leq x < 260.5$	7	2
261 – 265	𝍸	5	$260.5 \leq x < 265.5$	5	1

(b)

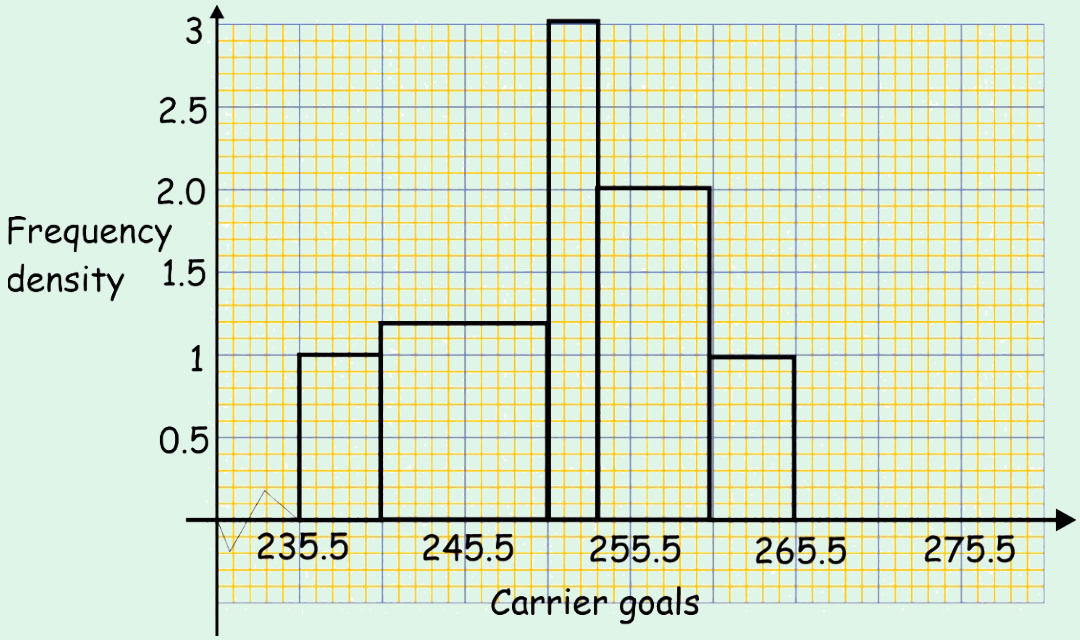


Figure 30: A histogram showing career goals

23.

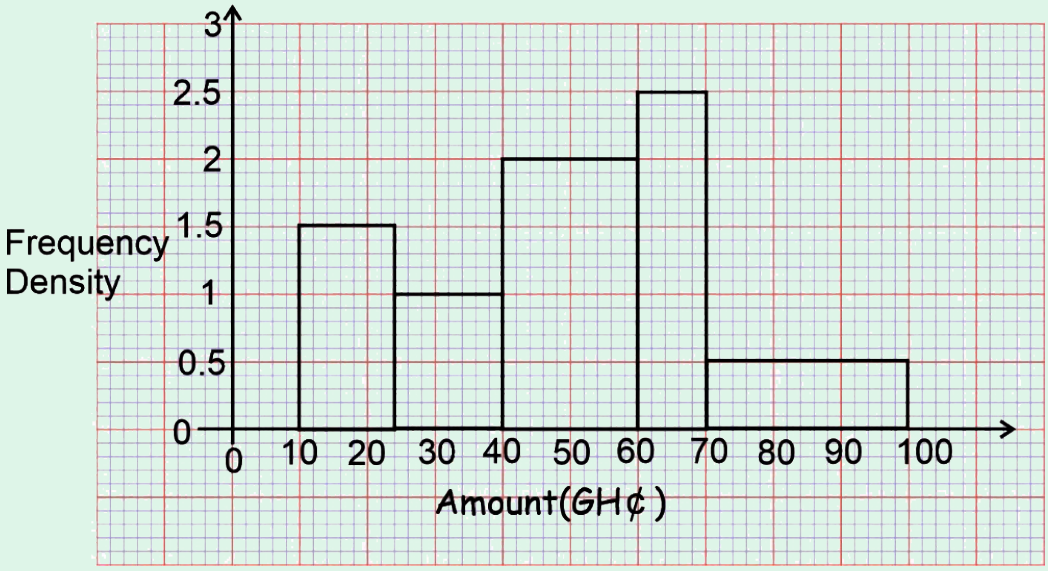


Figure 31: A histogram showing amount spent

24. (a)

Speed (km/h)	Class width	F	Frequency density
20 – 30	10	15	1.5
30 – 40	10	12	1.2
40 – 56	16	48	3
56 – 70	14	28	2
70 – 76	6	15	1.5
76 – 90	14	14	1

(b) 75 cars

(c)

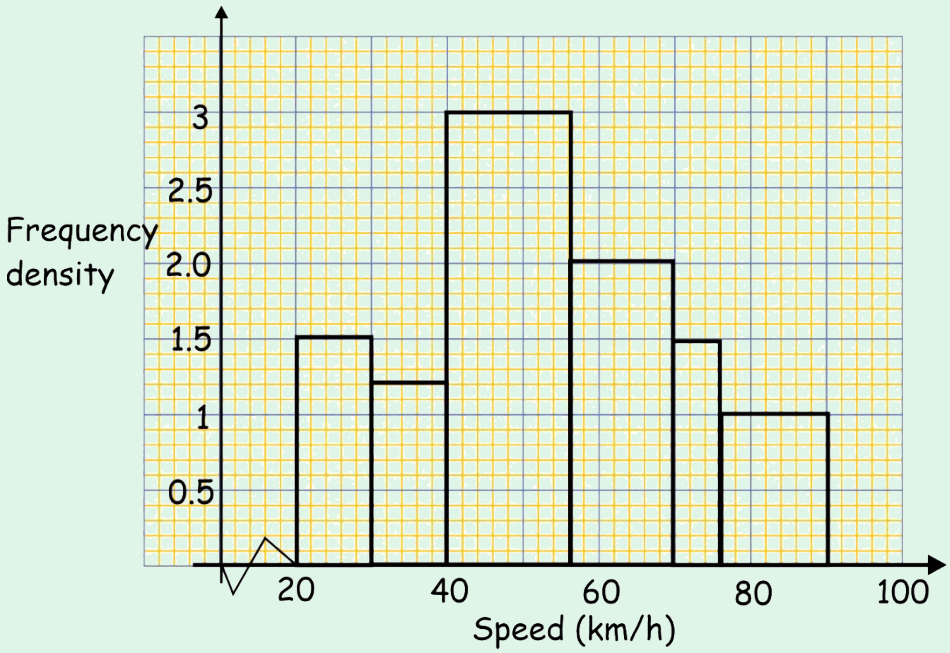


Figure 32: A histogram based on speed records of a policewoman

(d) GH¢ 11 600

25. (a)

Age (years)	Number of children
6-11	13
11-14	18
14-16	19
16-19	21
Total	71

(b) 12 students

26. (a)

Life expectancy (years)	Class Boundaries	F	C.F
1 – 5	$0.5 \leq x < 5.5$	6	6
6 – 10	$5.5 \leq x < 10.5$	4	10
11 – 15	$10.5 \leq x < 15.5$	14	24
16 – 20	$15.5 \leq x < 20.5$	18	42
21 – 25	$20.5 \leq x < 25.5$	15	57
26 – 30	$25.5 \leq x < 30.5$	11	68
31 – 35	$30.5 \leq x < 35.5$	8	76
36 – 40	$35.5 \leq x < 40.5$	4	80

(b)

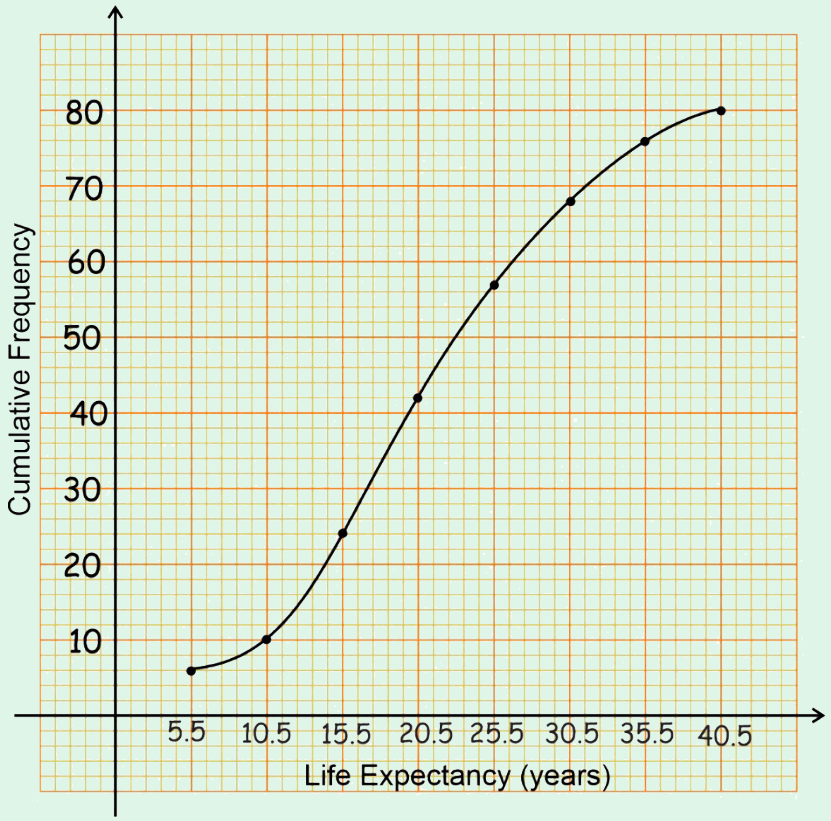


Figure 33: A cumulative frequency curve of animal life expectancy

27. (a)

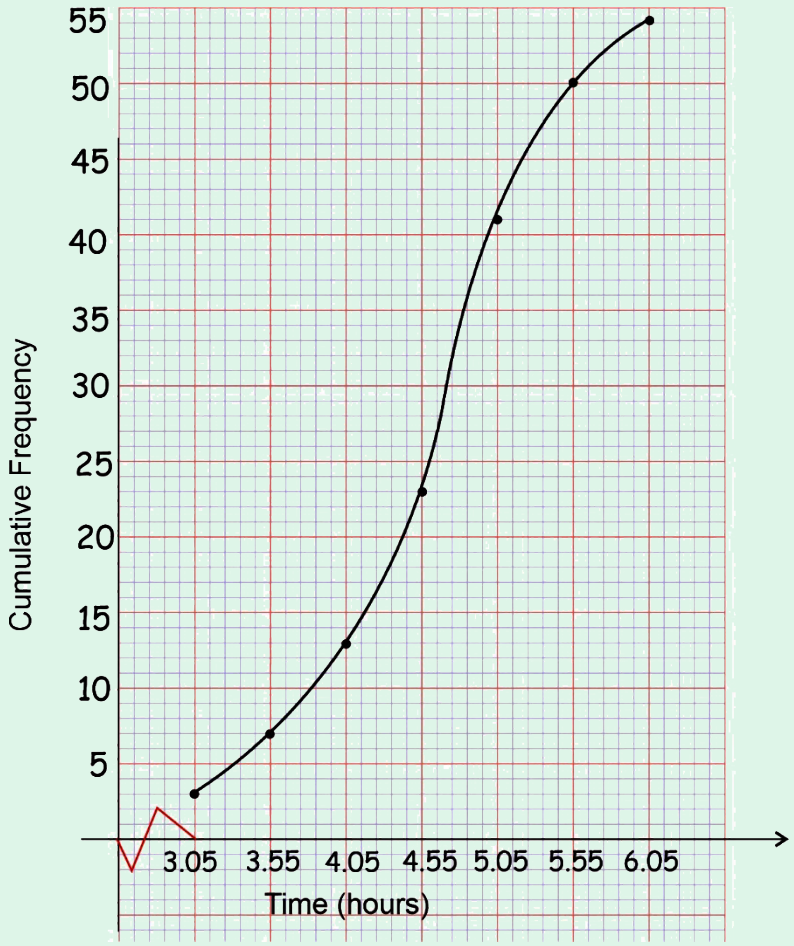


Figure 34: A cumulative frequency curve of the length of time to complete a challenge

(b) 4.39528. (a)

Class boundaries	Cumulative frequency	Frequency
1.5 – 2.5	3	3
2.5 – 3.5	5	2
3.5 – 4.5	8	3
4.5 – 5.5	13	5
5.5 – 6.5	20	7
6.5 – 7.5	33	13
7.5 – 8.5	41	8
8.5 – 9.5	45	4
9.5 – 10.5	47	2

(b) 7.0

(c) 6.4255

29. **(a)** Median = 26

(b) Lower Quartile = 22

(c) Upper Quartile = 35, so IQR = 13

30. IQR = 607 – 308.5 = 298.5, Standard Deviation = 181.899 (to 3dp)

31. **(i)** IQR = 10 – 8 = 2

(ii) Variance = 1.9316, SD = 1.3898

GLOSSARY

- **Data:** this refers to the raw facts, figures or measurements collected through observation, research or experiments, which can be analysed to derive meaningful information.
- **Discrete Data**: this consists of distinct, separate values that can only take specific numbers. These values are often counted (e.g., number of students in a class) and do not allow for fractions or decimals.
- **Continuous Data**: such data can take any value within a given range and is typically measured rather than counted. It includes data that can have an infinite number of possible values, such as height, weight and time.
- **Statistics**: this is the branch of mathematics that deals with the collection, organising, analysis, interpretation and presentation of numerical data. It involves using data to make informed decisions and predictions.
- **Sampling**: this is the process of selecting a subset of individuals, items or data points from a larger population for the purpose of making inferences about the entire population.
- **Sample Size**: this refers to the number of observations or data points selected from a population in a study or survey. It plays a crucial role in determining the accuracy and reliability of statistical results.
- **Grouped Data**: this is data that has been organised into categories or intervals (called classes) for easier analysis. The data within each group are combined or summarised.
- **Ungrouped Data**: implies raw data that has not been classified or organised into categories. It is presented in its original form, as individual values.
- **Frequency**: this refers to the number of times a specific value or category appears in a dataset. It is often displayed in a frequency table to show how often each value occurs.
- **Dispersion:** tells you how spread out the numbers are in a group. It shows whether the numbers are close together or far apart.

EXTENDED READING

Follow the steps outlined to practice how to find the mean, standard deviation and variance using Microsoft Excel.

Steps:

1. Open Microsoft Excel.
2. In any column (e.g., column A), enter your set of observations, one value per cell.

   For example, enter data in cells **A1** to **A10** if you have 10 observations.
3. Click on an empty cell where you want the **mean** to appear (e.g., cell **B1**).
4. Type the formula: =AVERAGE(A1:A10)

   Replace A1/A10 with the actual range of your data
5. Press Enter and the mean of your data will be displayed in the selected cell.
6. Click on an empty cell where you want the **standard deviation** to appear (e.g., cell **B2**).
7. Type the formula for **sample standard deviation**:

   =STDEV.S(A1:A10)

   If you're working with **population data**, use:

   =STDEV.P(A1:A10)

   Replace A1/A10 with the actual range of your data
8. Press Enter and the standard deviation will be displayed in the selected cell.
9. Type the formula for **sample variance**:

   =VAR.S(A1:A10)

   If you're working with **population data**, use:

   =VAR.P(A1:A10)

   Replace A1 with the actual range of your data
10. Press Enter and the variance will be displayed in the selected cell.

REFERENCES

- J B Ofosu 'O' Level Additional Mathematics Study guide and worked examples, Volume1
- Tuttuh-Adegun, M.R & Adegoke, D. G. New (2011) *Further mathematics project* Bounty Press Limited, Ibadan
- Otoo, D.K. (2005) *Concise elective mathematics*, ISBN 9988-0-2259-X
- Backhouse, J.K & Houldsworth, S.P.T. (1985) *Pure mathematics 1*
- Baffour, A. (2018). *Elective mathematics for schools and colleges*. Baffour Ba Series. ISBN: P0002417952

ACKNOWLEDGEMENTS

Ghana Education Service (GES)

List of Contributors

Name	Institution
Yaw Efa	Ghana National College
Benedicta Ama Yekua Etuaful	Ogyeedom SHTS
Isaac Buabeng	Ghana Education Service, Accra Metropolitan
Mpeniasah Kwasi Christopher	Three-Town SHS