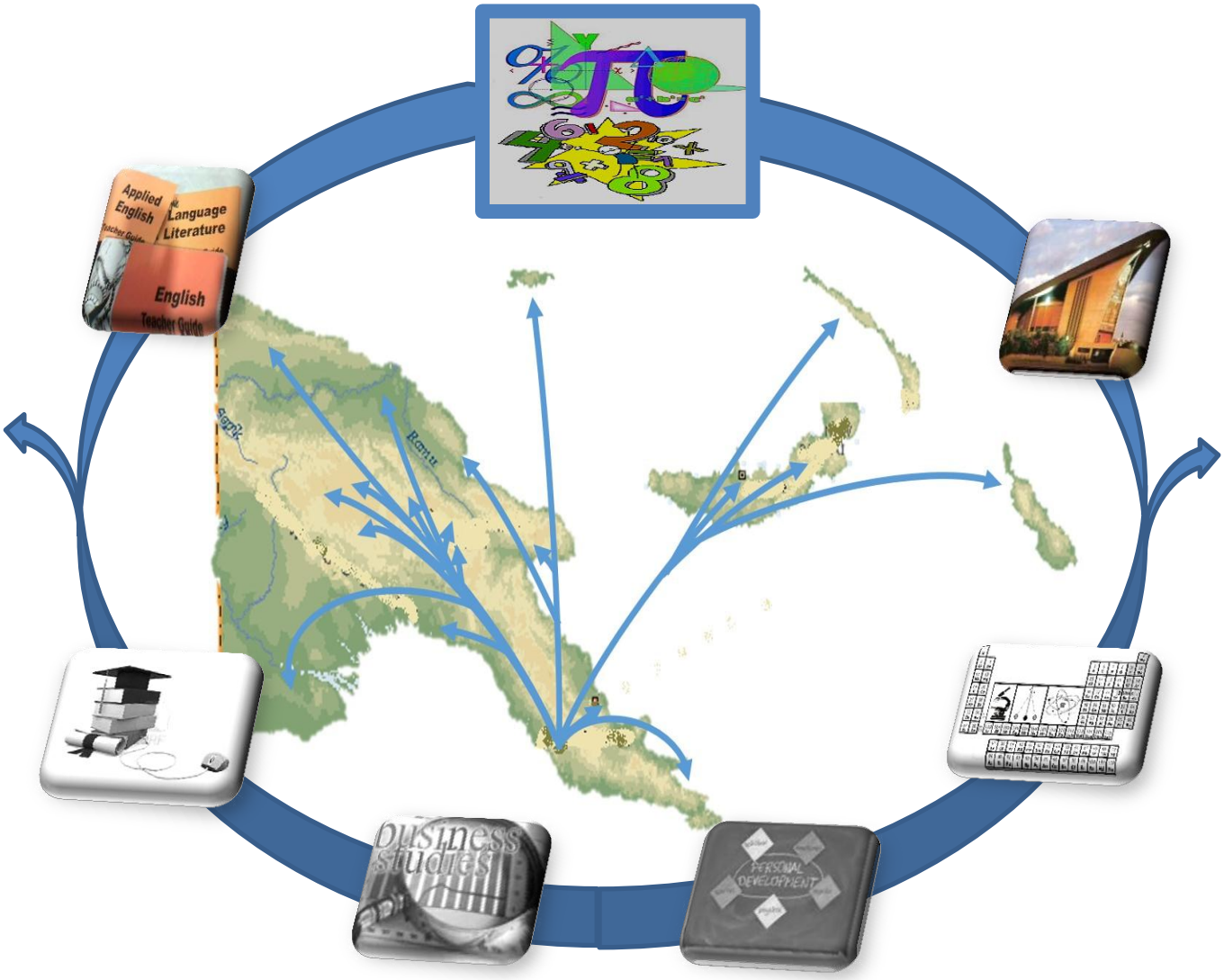
DEPARTMENT OF EDUCATION

GRADE 11 GENERAL MATHEMATICS

11.3: STATISTICS



FODE DISTANCE LEARNING

PUBLISHED BY FLEXIBLE OPEN AND DISTANCE EDUCATION
FOR THE DEPARTMENT OF EDUCATION
PAPUA NEW GUINEA
2017

GRADE 11

GENERAL MATHEMATICS

MODULE 3

Statistics

ACKNOWLEDGEMENT

We acknowledge the contributions of all Secondary Teachers who in one way or another have helped to develop this Course.

Our profound gratitude goes to the former Principal of FODE, Mr. Demas Tongogo for leading FODE team towards this great achievement. Special thanks to the Staff of the Mathematics Department of FODE who played an active role in coordinating writing workshops, outsourcing lesson writing and editing processes, involving selected teachers of Central Province and NCD.

We also acknowledge the professional guidance provided by Curriculum and Development Assessment Division throughout the processes of writing, and the services given by member of the Mathematics Review and Academic Committees.

The development of this book was Co-funded by GoPNG and World Bank.

DIANA TEIT AKIS

PRINCIPAL

Flexible Open and Distance Education
Papua New Guinea

Published in 2017

ISBN 978 9980 89 347 5
National Library Services of Papua New Guinea

Compiled and finalised by: Mathematics Department-FODE

Printed by the Flexible, Open and Distance Education

CONTENTS

:

SECRETARY'S MESSAGE

Achieving a better future by individuals students, their families, communities or the nation as a whole, depends on the curriculum and the way it is delivered.

This course is part and parcel of the new reformed curriculum – the Outcome Base Education (OBE). Its learning outcomes are student centred and written in terms that allow them to be demonstrated, assessed and measured.

It maintains the rationale, goals, aims and principles of the National OBE Curriculum and identifies the knowledge, skills, attitudes and values that students should achieve. This is a provision of Flexible, Open and Distance Education as an alternative pathway of formal education.

The Course promotes Papua New Guinea values and beliefs which are found in our constitution, Government policies and reports. It is developed in line with the National Education Plan (2005 – 2014) and addresses an increase in the number of school leavers which has been coupled with a limited access to secondary and higher educational institutions.

Flexible, Open and Distance Education is guided by the Department of Education's Mission which is fivefold;

- to facilitate and promote integral development of every individual
- to develop and encourage an education system which satisfies the requirements of Papua New Guinea and its people
- to establish, preserve, and improve standards of education throughout Papua New Guinea
- to make the benefits of such education available as widely as possible to all of the people
- to make education accessible to the physically, mentally and socially handicapped as well as to those who are educationally disadvantaged

The College is enhanced to provide alternative and comparable path ways for students and adults to complete their education, through one system, many path ways and same learning outcomes.

It is our vision that Papua New Guineans harness all appropriate and affordable technologies to pursue this program.

I commend all those teachers, curriculum writers and instructional designers, who have contributed so much in developing this course.

Dr. UKE KOMBRA, PhD
Secretary for Education

MODULE INTRODUCTION

Statistics is the science of collection, presentation, analysis and interpretation of data. The study of statistics is important because we frequently organize data numerically and make conclusions and use that information to influence our decisions.

We use statistics in making predictions, surveys and research. Education today is research based and teachers require their students to make a thesis. In this sense, statistics plays a vital role in education. We need to process statistical information precisely and accurately to function as knowledgeable citizens of the society.

Along with the advancement of technology that directly influences our life, the need to study statistics; statistics has grown enormously in the past five decades. The information that are presented through newspapers, magazines, televisions and the internet are all product of statistics. Through statistics ideas can easily be analyzed and answers to questions and queries can easily be defined or answered.

The topics in this module include:

Topic 1 EXPLORING DATA

This topic gets us to explore types of numerical data and data presentations; the frequency distribution that enables us to classify data to make comparison. We will restrict ourselves to discuss the assumed normal distribution among other distributions. We will also discusses the frequency histogram and polygon, stem-and-leaf-plot, cumulative frequency distribution and relative frequency of samples or population data of normal distribution.

Topic 2 MEASURES OF CENTRAL TENDENCY

This topic covers the four moments of statistics in mean, standard deviation, skewness and kurtosis . It starts with the three central tendencies in mean, mode and median. It then illustrates the relation between the standard deviation in skewed and normal distribution. Skewness illustrates the lack of symmetry about a mean and measures the relative size of tails in a given distribution. While kurtosis is a measure of combined sizes of tails.

,

Topic 3 MEASURES OF SPREAD OR DISPERSION

This topic gets us to discusses the range, semi and interquartile range, average and standard deviation of ungrouped data and the variance. We will explore and explain how a normal distribution of values are located about an average (mean or median).

The key to good data analysis is maintaining a balance between getting a good distribution fit and preserving ease of estimation, with ultimate aim that the analysis leads to sound decision. We must be aware that most data do not meet the criteria needed for distribution to fit due to asymmetry in data.

Student Learning Outcomes

On successful completion of this module, students will be able to:

- demonstrate the application of statistical knowledge and probability to communicate, justify, predict and critically analyze findings and draw conclusions;
- communicate mathematical processes and results;
- undertake mathematical tasks individually and/or cooperatively in planning, organizing, and carry out mathematical activities;
- construct stem-and-leaf plots;
- plot frequency polygons and histograms;
- tabulate and plot cumulative frequency distribution;
- state whether the data is skewed to the left or right of the mean;
- interpolate and extrapolate and/or calculate percentile, quartile and interquartile ranges;
- calculate range, interquartile range of any given data; and
- list and calculate the types of deviation as mean, variance and standard deviation.

Time Frame

This module should be completed within 10 weeks.

If you set an average of 3 hours per day, you should be able to complete the module comfortably by the end of the assigned week.

Try to do all the learning activities and compare your answers with the ones provided at the end of the module. If you do not get a particular exercise right in the first attempt, you should not get discouraged but instead, go back and attempt it again. If you still do not get it right after several attempts then you should seek help from your friend or even your tutor. Do not pass any question without solving it first.

11.3. 1: EXPLORING DATA

Data are facts or a set of information or observations that we gather either through a survey or an experiment. Sometimes statistics is misused when these data are used by some advertisers and campaign managers to promote the image or brand of their product in advertisements.

Thus, it is really important that we should be informed about the processes of collection, analysis and interpretation of data. With the knowledge of analysis and interpretation of data we can make independent decision on whether or not to agree with those who manipulate data to mislead us.

11.3.1.1 Statistical Data

In our daily activities, we encounter a lot of sorting and organizing objects, data, or things. These are just few of the activities of doing Statistics.

The data that we gather are assembled, classified and tabulated so as to present significant information about the nature of the gathered data. These are then analyzed and valid conclusions are drawn from the analyses for making informed decisions.

Data may be classified into two major types.

> **Qualitative Data** are information that are descriptive in nature. It refers to attributes like the information that we write in a curriculum vitae or bio-data.
>
> **Quantitative Data** are numerical information obtained through information or counting.

Examples

Qualitative Data
gender, religion, citizenship, educational attainment, etc.

Quantitative Data
age, IQ scores, height, weight, etc.

Qualitative data is also referred to as **categorical data**. Categorical data can be either ordinal or nominal data. Ordinal data are as such as …., and examples of nominal data are ….

And quantitative data is referred to as **numerical data**. The numerical data gathered about a sample can either be **discrete** or **continuous** data. Examples of discrete data are the number

of students in a class, the number of chairs in a classroom, etc. Examples of continuous data are age, length, weight, height, volume, time, etc.

Discrete Data are values that are the result of counting.

Continuous Data are values obtained through the process of measurement.

Data Collection

Collecting data can be done as follows:

1. Direct or interview method
2. Indirect or questionnaire method
3. Registration method
4. Experimental method

Example 1 Consider the published survey result below and try to answer the questions that follow.

ON-LINE SURVEY

Are you in favour of legalizing divorce?

YES 53%

NO 47%

These results are taken from a survey with 300 respondents.

- How were the respondents to the survey selected?
- How was the data collected?

The survey report does not say how the respondents were selected. The respondents could be from the same area, same town, and does not cover a broader area. Likewise, it does not say how the data was selected. When reading survey reports it is important to consider the following before drawing conclusions.

1. the source of the statistical information
2. the procedure on selecting the respondents
3. the process of collecting data
4. deliberate omissions of information or oversight.

You may notice that many things can go wrong in the compilation of statistical data which can make the survey result useless and or misleading. Thus, **precision** and **accuracy** in **gathering and presenting data** should be observed.

Learning Activity 11.3.1.1

20 minutes

A. Classify the following as quantitative or qualitative data:

1. Colour of the eye _____
2. Score in a Statistics _____
3. Temperature _____
4. Geographic location _____
5. Nationality _____
6. Gender _____
7. Length of objects _____
8. Mass of objects _____
9. Time interval _____
10. Racial discrimination _____

B. Identify each of the following types of data as either continuous or discrete.

1. Area of a land _____
2. Books in a library _____
3. Weight of students _____
4. Dimensions of a table _____
5. Number of enrolees in a certain school _____
6. Amount of water _____
7. Room Temperature _____
8. Speed of a car _____

11.3.1.2 Frequency Distribution

Data gathered can be presented in three ways.

Textual form is used when the data to be presented are few.

Example

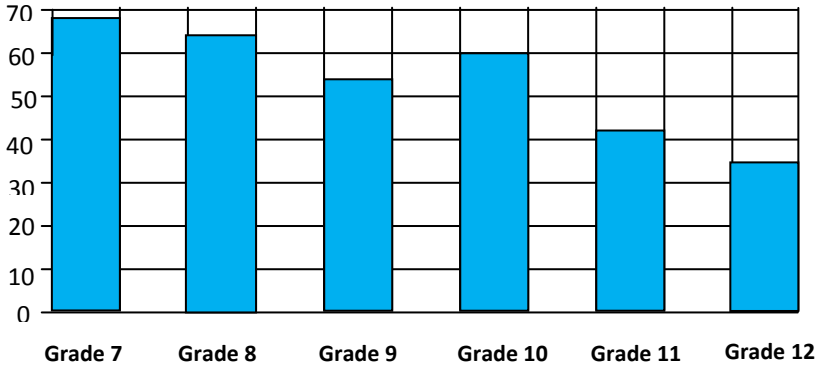
Number of Enrolees in a remote Public High School

Grade 7	-	68
Grade 8	-	65
Grade 9	-	57
Grade 10	-	61
Grade 11	-	42
Grade 12	-	36
Total	=	329

The second is Graphs. Graphs can be Line Graph, Pictograph, Bar Graph, Column Graph and the Circle or Pie Graph.

Graphical Form is used if data are presented through graphs.

Example **Number of Enrolees in a remote Public High School**



The third type , the tabular form is when the data is presented in a table.

Tabular Form is more practical and convenient to use and the data are usually arranged in columns or rows.

Grade	Class Size
7	68
8	65
9	57
10	61
11	42
12	36

Data can be classified as qualitative or quantitative or discrete or continuous data. In data presentation the information gathered, whether qualitative or quantitative or, discrete or continuous can be presented as either grouped data or ungrouped data.

Raw data are data gathered and listed in no particular order. These data are not organized numerically. Otherwise, these data are presented in ungrouped or grouped frequency distribution.

Raw Data of Ungrouped data : 4,7,5,9,9,8,7,6,8,7,7

Distribution of Ungrouped : 4.5.6.7.7.7.8.8.9

Frequency Distribution is a way of classifying statistical data arranged in order of size that allows comparisons of the results in each category.

Frequency Distribution is a tabular arrangement of the data by categories showing the frequency of occurrence of values and class marks or midpoints. It has a class frequency containing the number of class intervals. With large data, tally is necessary to aid attain accurate frequency.

Ungrouped Frequency Distribution: Scores are arranged individually in order of size, for example from lowest to highest or ascending order.

Grouped Frequency Distribution: This is used if the number of scores is big and when comparisons are made between several groups of continuous data.

In a Grouped Distribution we use the **mid-point** of a group or class to compute the mean.

Mid-point = ½ (lower limit + upper limit)

Rules in forming Grouped Frequency Distribution.

1. Get the difference between the largest and smallest number (or value) in the raw data.

Thus, we determine the range using this formula:

Range = Highest Score – Lowest Score

2. Solve for the class interval size by dividing the range by the expected number of classes (ideally 8 to 12 classes) then round off the result so that the class interval size is a whole number.

Note: As a general rule, the class size is preferred to be odd so that the midpoint will be a whole number.

3. Determine the class limits. There must be enough classes to include the highest score and lowest score. To do this tabulation, start each class with a multiple of the class interval.

4. Find the number of observations falling into each class interval. To make the tabulation, the table should have at least two columns. The first column shows the classes usually in descending order from top to bottom. While the second column shows the frequency which are the number of observations for each class.

Example 1 The following are the results of the 50-item test in a Grade 11 class with 50 students. Construct a frequency distribution from the given raw data.

28	34	26	41	45	36	33	41	41	39
21	30	35	***19***	29	39	47	47	42	44
42	45	41	46	37	35	46	47	24	46
43	43	41	45	39	47	42	41	38	47
41	40	35	36	42	47	44	40	37	47

Solution

Step 1: Range = Highest Score - Lowest Score
Range = 47 – 19
Range = 28

Step 2: Divide the range by the expected number of classes (say 10)
28 ÷ 10 = 2.8
Class interval size (*i*) = 3 Rounded off to the nearest whole number

Step 3: Determine the class limits per class (first column usually in descending order.

Note: Since the highest score is 47, we start the upper limit from 48 since it is divisible by 3.

Step 4: Use the class interval to form classes and Tabulate

Descending Order

Classes	*f*	*Midpoint* (X)
46 – 48	10	47
43 – 45	7	44
40 – 42	13	41
37 – 39	6	38
34 – 36	6	35
31 – 33	1	32
28 – 30	3	29
25 – 27	1	26
22 – 24	1	23
19 – 21	2	20

Ascending Order

Classes	*f*	*Midpoint* (X)
19 - 21	2	20
22 – 24	1	23
25 – 27	1	26
28 – 30	3	29
31 – 33	1	32
34 – 36	6	35
37 – 39	6	38
40 – 42	13	41
43 – 45	7	44
46 - 48	10	47

You may notice that we produced ten (10) class intervals since we divide the range by 10. Likewise, the midpoint or class centre denoted by **X** was included in the third column which is obtained by getting the average of each class limits.

Example What is the mid-point of the Class 40 - 42?

Solution

Class centre (X) = ½ (lower class limit + upper class limit)
= ½ (40 + 42)
= ½ x 82
= 41

The class frequency tells how many items or scores fall into each class.

Example What is the frequency of the Class 40 - 42

Solution

f = 13

The smaller values in each class, say 46 in the first class (46-48), are called the **lower class limits** and the larger values are called the **upper class limits**.

To avoid gaps in the continuous number scale (data), it is usual to subtract 0.5 from each lower limit to refer to the **lower class boundary** and add 0.5 to each upper limit for the **upper class boundary.**

If the set of data is given correct to one decimal place, we add or subtract 0.05 to obtain upper and lower class boundaries. Likewise, suppose the data is given to the nearest ten, we add or subtract 5 to obtain upper and lower class boundaries.

The difference of the class boundaries gives more accurate class size than the difference of class limits.

Example Using the frequency distribution table on the results of the 50-item test in a Grade 11 Class with 50 students, for the class interval 28 – 30, find:

a. class mark
b. lower class limit
c. upper class limit
d. lower class boundary
e. Upper class boundary

Solution

Class Interval	Class Mark (X)	Lower Class Limit	Upper Class Limit	Lower Class Boundary	Upper Class Boundary
28 – 30	29	28	30	27.5	30.5

Now to be able to analyze the scores, we need to find fX of the Frequency distribution table. The fX is the product of the frequency and the class centre.

We also need to find N or ∑f (number of scores) and ∑fX (sum of products).

Study the frequency distribution table.

Classes	*f*	*Midpoint* (X)	fX	Lower Class Boundary	Upper Class Boundary
19 - 21	2	20	40	18.5	21.5
22 – 24	1	23	23	21.5	24.5
25 – 27	1	26	26	24.5	27.5
28 – 30	3	29	47	27.5	30.5
31 – 33	1	32	32	30.5	33.5
34 – 36	6	35	150	33.5	36.5
37 – 39	6	38	228	36.5	39.5
40 – 42	13	41	533	39.5	42.5
43 – 45	7	44	308	42.5	45.5
46 - 48	10	47	470	45.5	48.5
	∑f = 50		∑fX = 1857		

From this table, we can read off or calculate mean, mode and median now that we have computed sum of frequencies, and sum of all scores based on the sum – product of frequency and mid-point value.

We are now in position to interpret queries directly related to the table, or we can interpret queries in relation to the table and the measures of central tendencies.

11.3.1.3 The Histogram and the Frequency Polygon

You have just learned how to construct frequency distribution as a way of presenting the gathered data in tabular form. In this lesson, you will learn how to present data in graphical form.

A **bar graph** uses bars of different lengths (depending on the frequency of each category) and of equal widths. They are drawn vertically or horizontally with equal distance from each other. A bar graph is used in presenting ungrouped data.

Example Teacher Anne asked her Grade 11 students to answer the questionnaire below.

SURVEY ON PREFERRED KIND OF MOVIE

What is your favourite kind of movie?
Check only one.

☐ Fantasy ☐ Horror ☐ Action

☐ Drama ☐ Love Story

Teacher Anne tallies the results of the survey and presents the data in a frequency table.

Preferred Kind of Movie	Tally	Number of Students
Fantasy	𝍸 - 𝍸 – I	11
Horror	𝍸 – IIII	9
Action	𝍸 – 𝍸 – 𝍸	15
Drama	𝍸 – 𝍸 – 𝍸 – 𝍸 – II	22
Love Story	𝍸 – 𝍸 – 𝍸 - II	17
Total		**74**

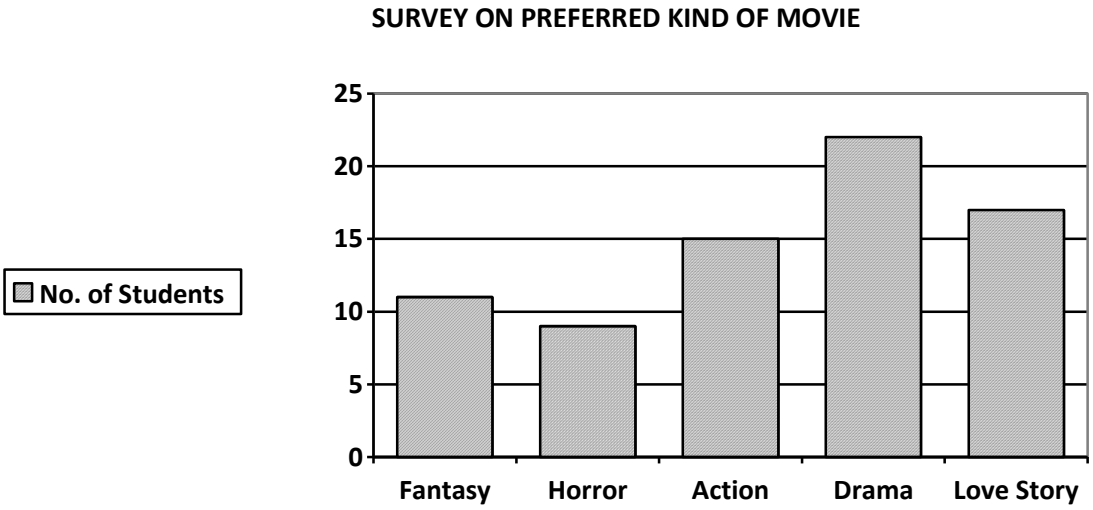
Construct a bar graph using the data presented in the frequency table and answer the following:

a. How many more students prefer drama than action?

b. Which kind of movie is least favoured?

c. How many students were surveyed in all?

Solution

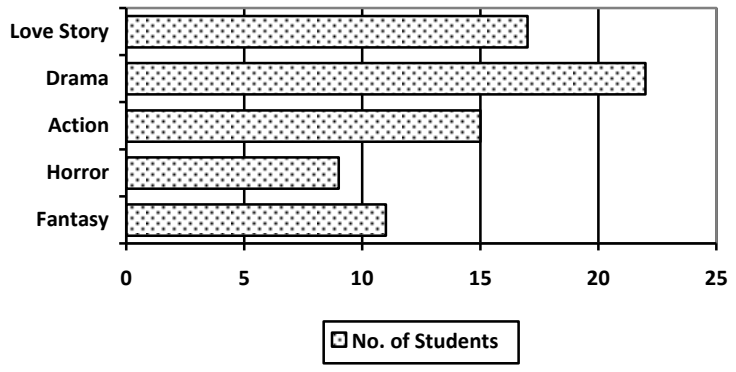


a. 22 – 15 = 7 Seven (7) students prefer drama than action.

b. Horror is the least favoured kind of movie.

c. 11 + 9 + 15 + 22 + 17 = 74
74 students were surveyed about their preferred kind of movie.

Note: Above is a column graph. The Bar Graph can also be presented where bars are drawn horizontally.



Histogram

In a histogram, the bars are always adjacent and vertically drawn. Histograms have no gaps because their base represents a continuous range of values and the width of each bar is based on the size of the interval it represents.

A **histogram** is a bar graph that shows the continuous ungrouped or grouped data.

Example Construct a histogram of the results of the surveyed age of 50 people.

25	36	48	50	55	60	70	76	40	53
34	52	63	41	29	45	53	71	31	33
83	26	39	62	53	21	74	37	28	75
45	56	71	38	64	55	32	21	34	80
23	47	63	39	48	27	58	67	49	52

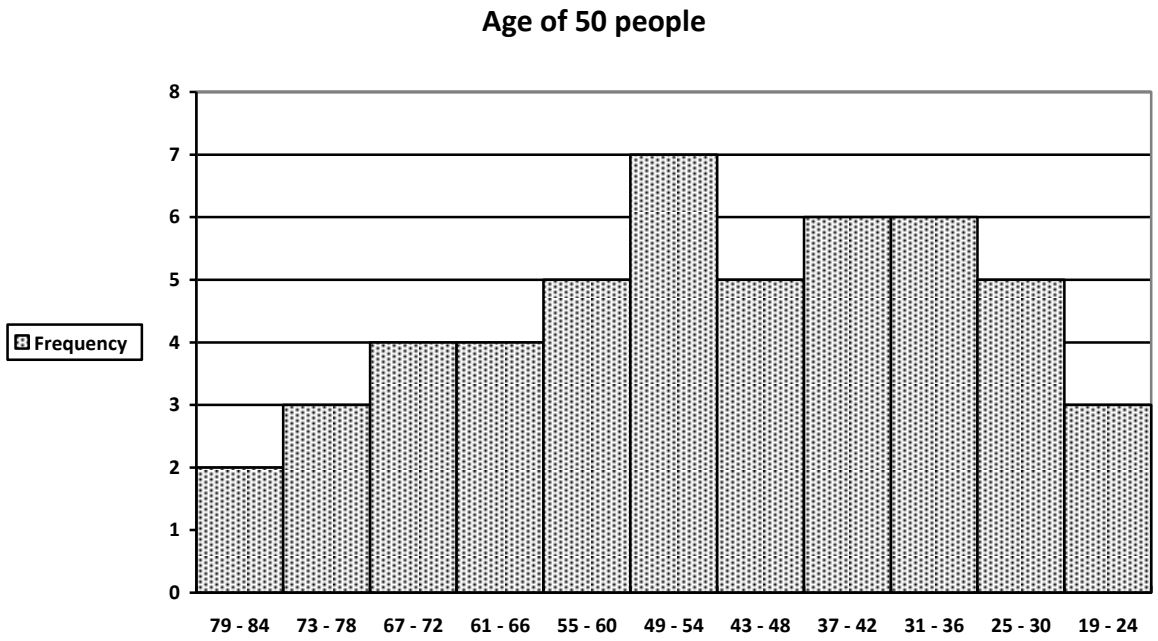
Solution

Divide the range by 10 and begin by constructing a frequency distribution table.

Class Intervals	Frequency
79 - 84	2
73 – 78	3
67 – 72	4
61 – 66	4
55 – 60	5
49 – 54	7
43 – 48	5
37 – 42	6
31 – 36	6
25 – 30	5
19 – 24	3
	N = 50

In constructing a histogram, it is important to write the title of the graph and labels. The vertical data is labelled as frequency while the horizontal data is labelled as the age of respondents.

Revisiting the definition of a histogram, instead of writing the class intervals on the horizontal, the class boundaries must be considered instead because the ages represents a continuous range of values.



Another form of presenting data is through line graph (or frequency polygon for grouped data).

A **line graph** is used when we want to show the falling and rising trend of a set of data over a period of time. The vertical line indicates the frequency while the horizontal line shows the categories being considered. A line graph is used when the data to be presented are few (ungrouped data).

Example The table shows the number of magazines borrowed in the library last week.

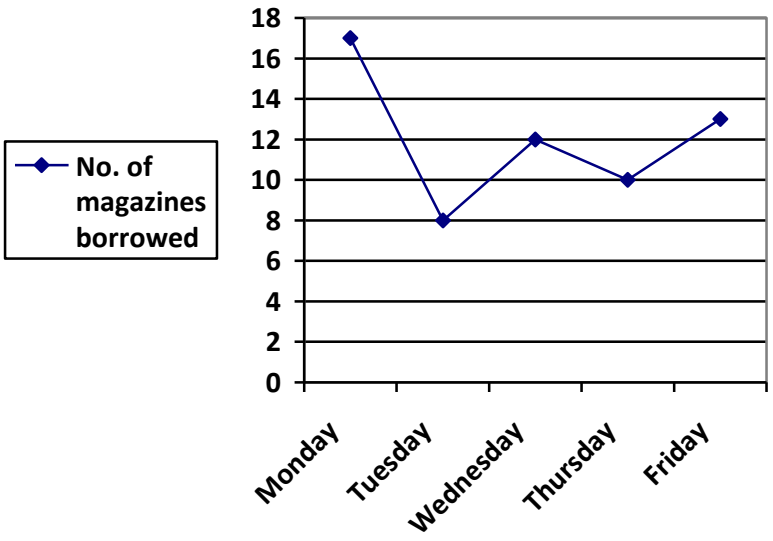
Monday	Tuesday	Wednesday	Thursday	Friday
17	8	12	10	13

Construct a line graph and answer the following:

a. How many magazines were borrowed on Friday?

b. What day had the most number of borrowed magazines?

c. How many magazines were borrowed in all last week?



Solution

a. 13 magazines were borrowed on Friday.

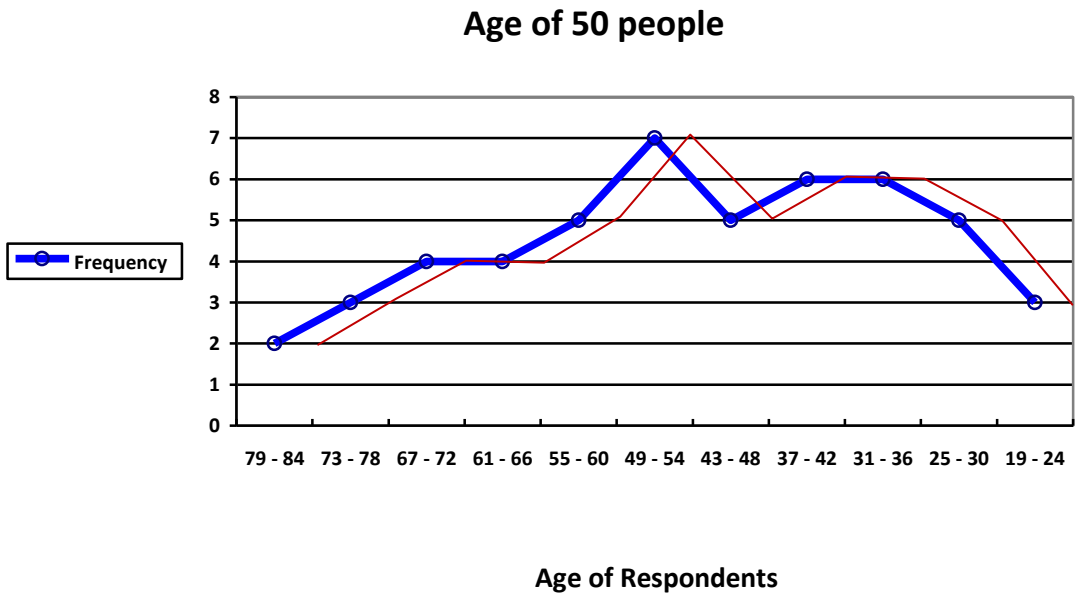
b. The most number of magazines borrowed last week was on Monday.

c. 60 magazines were borrowed in all last week.

A **frequency polygon** is a line graph of a class frequency plotted against mid-points of class mark or lower class boundaries or upper class boundaries.

Example Construct a frequency polygon using the frequency table on page 12.

Solution

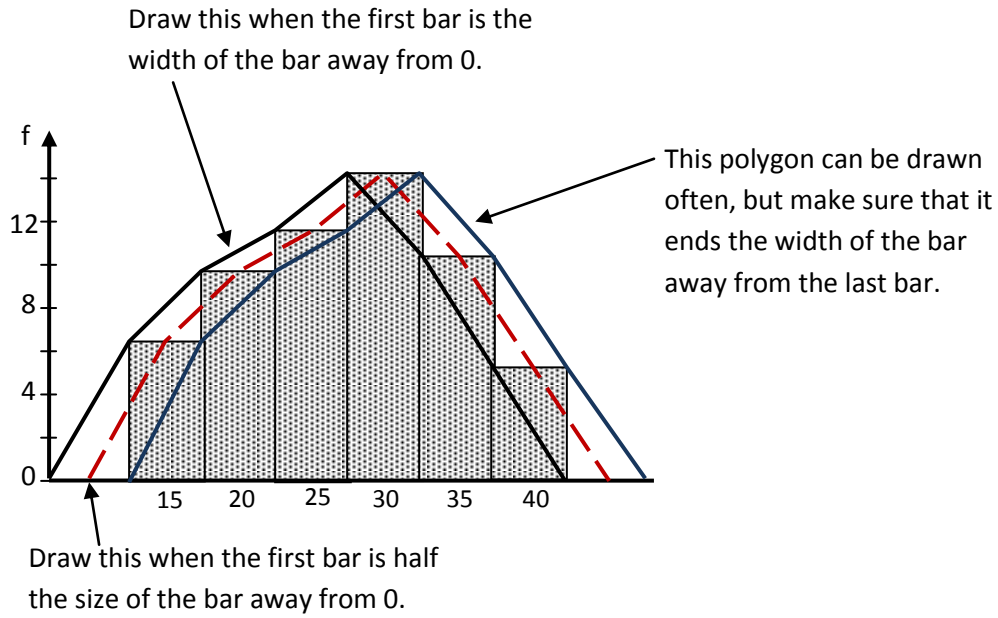


The three polygons represent the same data. The one in dashes was drawn by joining mid-points and the other one was drawn by joining the upper class boundaries, while the third was drawn joining the lower class boundaries as shown in the histogram below.

You can decide and chose one among the three different ways of constructing a polygon. But take note of notes given below, so that the areas below polygon and the areas covered by all the bars together is the same.

There should always be a 0.5 cm space between the lower class boundary and the frequency axis. Or the space can be half the width of the bar. If the bar width is 2 cm, the space should be 1 cm wide between the axis and lower boundary of the first bar. The polygon starts from zero.

If the space is the same as width of bar, polygon can start from 0 and join lower boundaries which are top left hand corners.

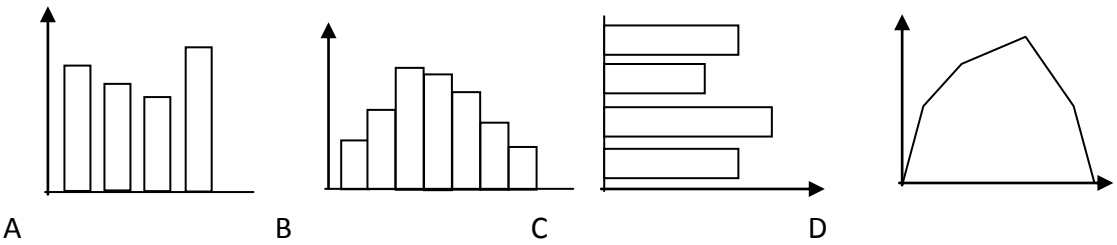


LEARNING ACTIVITY 11.3.1.2 – 11.3.1.3

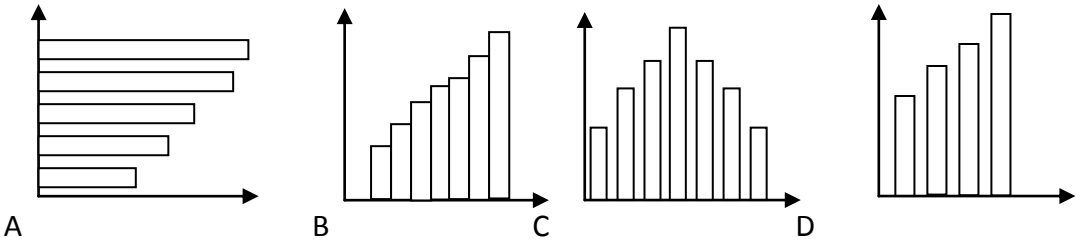
4 0 minutes

A. Multiple Choice – Circle the correct response.

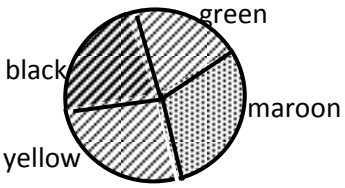
1. Which of the following represents a bar graph?



2. Which graph would NOT be appropriate to illustrate a discrete data?



3. What fraction of the pie graph represents the colour maroon?



A. $^1/_8$ B. ¼ C. $^1/_3$ D. ½

4. Another term for mid-point is

A. Class centre B. class limit C. class boundary D. class interval

5. Data arranged in ascending or descending order is referred to as a

A. Continuous data B. qualitative C. Frequency D. distribution

6. The upper class boundary of the class group 20 – 25 is

A. 6 B. 12. 5 C. 25 D. 25.5

7. The class interval of the class group 32 – 37 is

A. 4 B. 5 C. 6 D. 7

B. For questions 1 to 3, refer to the following information in question 1:

1. The following are the scores of 40 students in their 50-item exam in Mathematics. Create a frequency distribution table using the following data.

27	31	48	35	45	50	40	46
32	25	26	41	49	42	43	47
50	29	39	42	33	23	41	37
41	43	44	38	34	46	32	11
28	45	38	39	48	47	38	47

Note: Divide the range by 10.The lowest limit must start from a number that is divisible by the class size or class interval (*i*).

2. Construct a histogram and a frequency polygon using the frequency distribution table created from item #1.

3. Construct a cumulative frequency distribution table for the ungrouped data first by grouping them with class interval of five. Begin with the least score on the top down to the highest at the bottom.

11.3.1.4 Stem-and-Leaf Plots

Stem-and-leaf plots is a method of organizing data in which the **stem values** or the leading digit for each observation are listed in a column and the **leaf values** or the trailing digit for each observation are then listed beside the corresponding stem.

Example The grades in Statistics of 20 students are 86, 79, 85, 82, 81, 91, 78, 89, 80, 83, 76, 93, 88, 91, 77, 84, 81, 86, 81, and 78. Construct a stem-and-leaf plot diagram of the given grades.

Solution

Notice that the grades start with 7, 8 and 9. Let us split each number in two parts: a stem such as 7, 8 and 9, and a leaf such as 1, 2, 3, and so on.

Stem	Leaf
7	6, 7, 8, 8, 9
8	0, 1, 1, 1, 2, 3, 4, 5, 6, 6, 8, 9,
9	1, 1, 3

Be sure to write the leaves in ascending order and check that the number of leaves tallies with the total number of data collected.

You may notice from the stem-and-leaf plot diagram that the more common grade is 81, the highest grade is 93 and the lowest grade is 76.

The stem-and-leaf plot can also be used in comparing data between two independent groups. Refer to the example below and answer the questions that follow.

Example The following stem-and-leaf diagram represents the scores of the students in two classes for a common examination. Each section consists of 30 students.

Section A		Section B
Leaf	Stem	Leaf
1, 4, 5	4	2, 3, 6
1, 2, 4, 5, 6	5	0, 1, 3, 5, 6, 7, 7
0, 1, 2, 2, 2, 7	6	1, 2, 3, 4, 5, 7, 8
4, 5, 6, 6, 8, 8, 9	7	3, 4, 5, 5, 7
2, 3, 4, 5, 5, 8	8	0, 1, 3, 4, 5, 9
2, 4, 8	9	5, 6

a. Which section got the highest score? What is the highest score?

b. Which section obtained the lowest score? What is the lowest score?

c. Which section do you think did better in the examination?

Solution

a. Section A obtained the highest score of 98. The digit on the Stem column are on the tens place while the digits on the ones or units place that is why the highest score is 98.

b. Section A obtained the lowest score of 41. The lowest digit on the tens place is 4 and 1 is also the lowest digit on ones place.

c. Answers may vary. One way is to get the average score of each class.

To get the average of Section A, we divide the sum of all the scores by 30. Write the actual scores and find the sum.

41	44	45				
51	52	54	55	56		
60	61	62	62	62	67	
74	75	76	76	78	78	79
82	83	84	85	85	88	
92	94	98				

Sum = 2099 Average = Sum ÷ 30 **Average = 69.97**
The average score of Section A is 69.97.

To find the average of Section B, we do the same procedure.

42	43	46				
50	51	53	55	56	57	57
61	62	63	64	65	67	68
73	74	75	75	77		
80	81	83	84	85	89	
95	96					

Sum = 2027 **Average = 67.57**
The average of Section B is 67.57.

Conclusion. The computed average is a proof that Section A perform better in the examination since they obtained an average of 69.97 which is obviously higher than the obtained average of Section B.

Although the top scorer belongs to Section A, we cannot conclude that they perform better unless we compute the average. We may clarify this in the preceding lessons on averages.

LEARNING ACTIVITY 11.3.1.4

20 minutes

1. Thirty applicants were given a 40-item verbal ability test. Construct a stem-and-leaf plot

27	28	32	30	37
31	35	26	25	33
29	34	31	28	24
18	21	37	35	38
31	27	35	40	32
35	31	22	29	38
38	36	33	39	37
30	31	32	37	28

2. Draw a dot-plot for the data in question 1.

11.3.1.5 Cumulative Frequency Distribution (The Ogive)

The total frequency of all values less than the upper class boundary of a given class is called the **Cumulative Frequency** including the class interval.

Example The frequency distribution below shows the weight (in pounds) of Grade 11 students in a public high school.

Weight of 100 Grade 11 Students in a Public High School

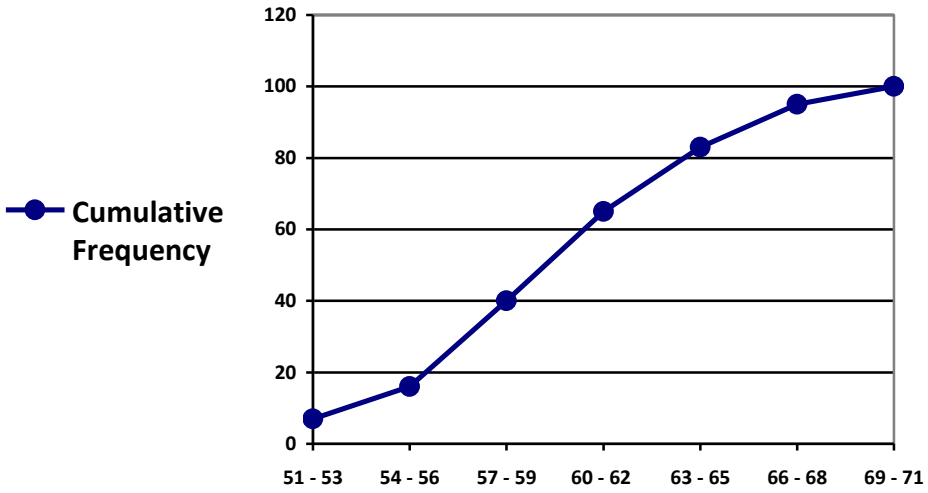
Weight (in kilograms)	Frequency (number of students)
69 – 71	5
66 – 68	12
63 – 65	18
60 – 62	25
57 – 59	24
54 – 56	9
51 – 53	7
TOTAL	N = 100

The cumulative frequency refers to the accumulation of scores or frequencies and it is obtained by adding to the present value the frequency of the previous class interval starting from the frequency of the class with the lowest value.

Weight (in kilograms)	Frequency (number of students)	Cumulative Frequency Less than (CV<)
69 – 71	5	100
66 – 68	12	95
63 – 65	18	83
60 – 62	25	65
57 – 59	24	40
54 – 56	9	16
51 – 53	7	7
TOTAL	N = 100	

An **OGIVE** is a graph that shows the accumulation of frequencies by class intervals arranged in a table. This is also called **Cumulative Frequency Polygon.**

Below is a graph showing the cumulative frequency less than on the weight of 100 Grade 11 students in a Public High School called **Cumulative Frequency Polygon** or **OGIVE.**

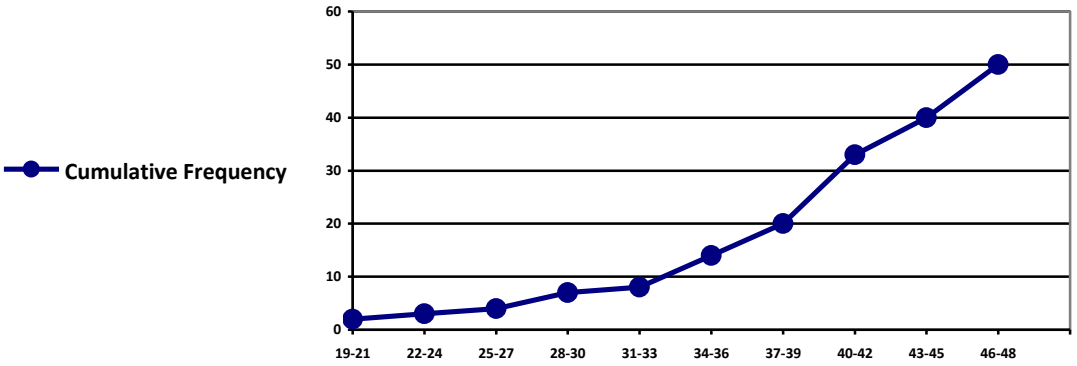


Example Construct a cumulative frequency polygon (OGIVE) using the table below.

Classes	*f*	Cf<
46 – 48	10	50
43 – 45	7	40
40 – 42	13	33
37 – 39	6	20
34 – 36	6	14
31 – 33	1	8
28 – 30	3	7
25 – 27	1	4
22 – 24	1	3
19 – 21	2	2

Solution

We construct a line graph labelling the vertical line as cumulative frequency less than (Cf<) and the horizontal line for the classes starting from the lowest class. We then plot points by aligning the actual Cf< and classes to show the trend of how the frequencies accumulates or increases.



11.3.1.6 The Relative Frequency

If cumulative frequency refers to the accumulation of frequency for each class, the relative frequency is the equivalent percentage of the frequency of each class. It can be obtained by dividing the frequency of each class by the total number of data then multiply the quotient by 100.

Example Complete the table below by finding the cumulative frequency less than and the relative frequency of each class.

Height of 45 Students in a PE Class

Classes (Height in cm)	Frequency (No. of students)	CF<	Relative Frequency (%)
175 - 179	6	60	10
170 – 174	8	54	13
165 – 169	10	46	17
160 – 164	15	36	25
155 – 159	12	21	20
150 – 154	9	9	15
	N = 60		100%

As explained in the previous lesson about Ogives, we add the frequencies starting from the lowest class (150-154) to complete the CF< column while the Relative Frequency column may be obtained by getting the percentage per class. This is obtained using the formula:

Relative Frequency = (frequency of each class ÷N) • 100%

Example Using the given example on page 20, complete the table by finding the relative frequency of each class.

Classes	*f*	Cf<
46 – 48	10	50
43 – 45	7	40
40 – 42	13	33
37 – 39	6	20
34 – 36	6	14
31 – 33	1	8
28 – 30	3	7
25 – 27	1	4
22 – 24	1	3
19 – 21	2	2

Solution

We may add another column for RF% using the formula.

Relative Frequency = (frequency of each class ÷N) • 100%

Classes	*f*	Cf<	Rf%
46 – 48	10	50	20%
43 – 45	7	40	14%
40 – 42	13	33	26%
37 – 39	6	20	12%
34 – 36	6	14	12%
31 – 33	1	8	2%
28 – 30	3	7	6%
25 – 27	1	4	2%
22 – 24	1	3	2%
19 – 21	2	2	4%
			100%

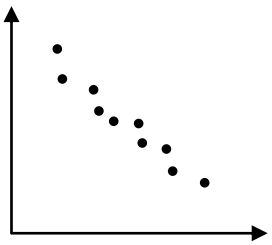
11.3.1.7 Scattergram and Correlation

When we compare two sets of random variables (bivariate data) to determine if there exist a linear relationship, we correlate between the sets of data. The graph used is a scattergram.

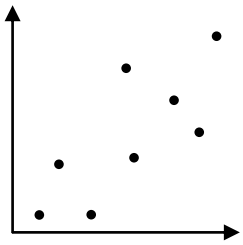
In a scattergram, the points are not collinear; we attempt to derive linear relationship so we can compare two sets of random variables.

Correlation is the extent of correspondence between the ordering of two random variables.

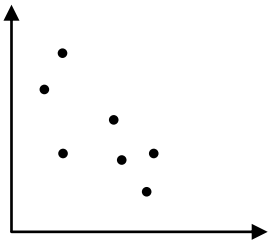
It is a positive correlation when each variable increases or decreases as the other does, negative or inverse correlation as one variable tends to increase while the other decreases. Correlation is high when points are clustered and seemingly linear. But correlation is low when points are scattered or spread.



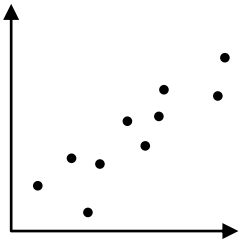
High negative correlation or strong negative correlation



Low positive correlation or weak positive correlation

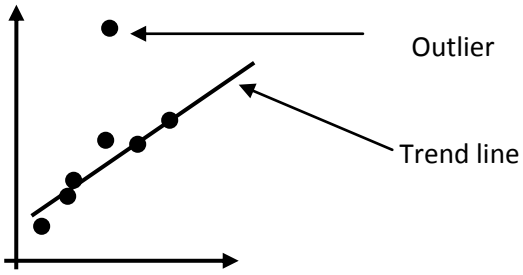


Low negative correlation or weak negative correlation



High positive correlation or strong positive correlation

When there is an outlier, it may affect the conclusion so omit the outlier.



Deriving Linear Equation of the Correlation

The equation enables us to approximate one result when the result is unknown, given that sufficient random variables are known. The linear equation is of the form y = mx + c, however in statistics it is often expressed as y = b + ax.

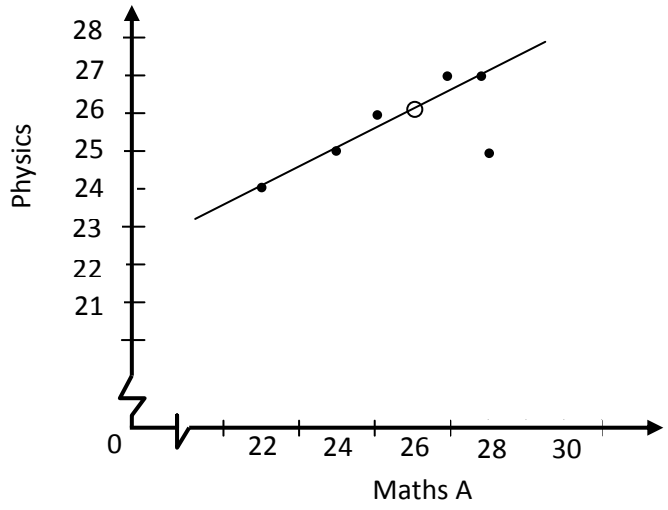
Steps

1. Plot the points
2. Find pivotal point (average of x, average of y)
3. Plot pivotal point $(\bar{x}, \bar{y})$
4. Rule a straight line (line-of-best-fit) through pivotal point and another point, having about equal points on either side of the line
5. Substitute into $y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$ where $m = \frac{y_2 - y_1}{x_2 - x_1}$ and (x_1, y_1) is the pivotal point and (x_2, y_2) is the other point the straight line runs through.
6. Simplify and express in the form y = b + ax where b is the y-intercept and a is the slope

Example Given below are the data of Mathematics A and Physics test results of a grade 12 student.

Maths A	24	26	22	28	25	27	28
Physics	25	26	24	25	26	27	27

a) Plot and derive the equation.



Pivotal Point $(\bar{x}, \bar{y})$ =(26, 26)

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$$

$$y - 26 = \frac{24 - 26}{22 - 26}(x - 24)$$

$$y = \frac{-2}{-4}(x - 24) + 26$$

$$y = \frac{1}{2}x - 12 + 26$$

$$y = \frac{1}{2}x + 14$$

Since Maths is x and Physics is y the equation becomes

$$P = \frac{1}{2}M + 14$$

$$P = 14 + \frac{1}{2}M$$

b) Describe the correlation between Maths A and Physics result of the student.

The relation is a strong positive correlation.

For learning experience, always use the pivotal and another point of your choice. But you must always ensue that when you rule a line-of-best-fit, there should be about equal points on either side of the line-of-best-fit.

Say suppose you use (24, 25) as other point along with the pivotal point (26, 26), then you will get, on substitution as

$$P - 26 = \frac{25 - 26}{24 - 26}(M - 26)$$

$$P - 26 = \frac{-1}{-2}(M - 26)$$

$$P - 26 = \frac{1}{2}M - 13$$

$$P = \frac{1}{2}M - 13 + 26$$

$$P = 13 + \frac{1}{2}M$$

The slope is the same, but the y-intercept changed slightly. That is anticipated. So always use pivotal point and another point on your line-of-best-fit to derive the equation and you will never go wrong.

You can use this skill to check correlation between the height and weight of your friends. Or correlation between time spent in studies and marks attained.

The more accurate method to derive equation of the line- of – best – fit (y = a + bx) is by least square regression technique. The **least squares regression** technique uses the distance of **the actual value** away from the **predicted value (** called the **residual)** and seeks to minimize the total of these distances**.**

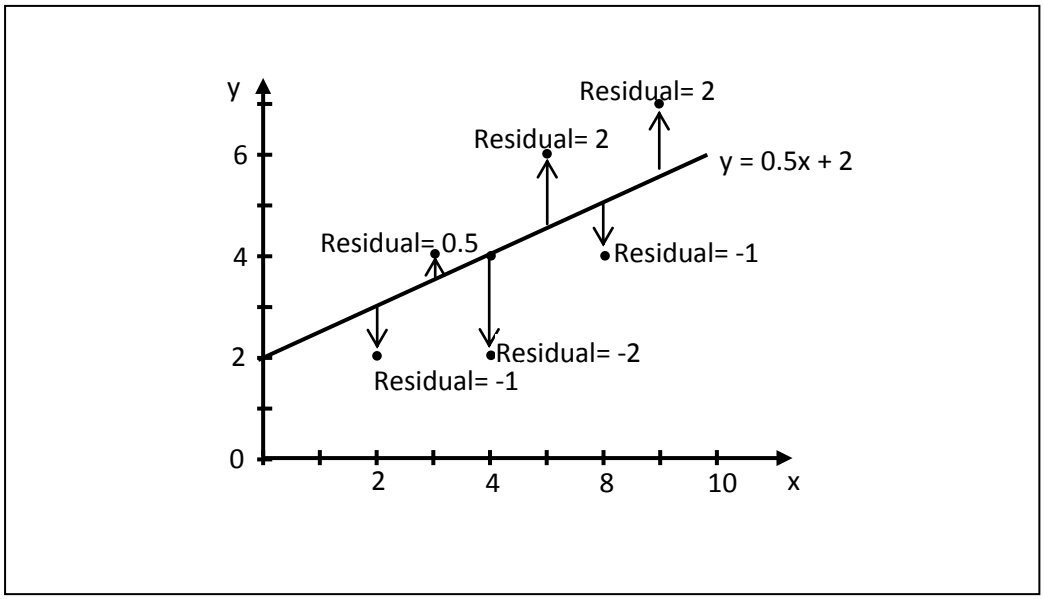
The method requires computation of **Pearson's correlation coefficient r** , the **means** and **standard deviations** of the bivariate data**.**

From the three values we can evaluate value of 'b', which is the slope of the trend line. Then we solve for 'a'

Pearson correlation coefficient r, or **product moment correlation coefficient** is a more accurate method to use to find correlation of bivariate data (two data sets). The formula is as given in the summary for your reference. The table reduces the task of solving for the 'r'.

We express as y = a + bx, a is the y-intercept and b is the gradient. Where $a = \bar{y} - b\bar{x}$, and $\bar{x}$ and $\bar{y}$ are respectively means of x and y values; and $b = \frac{rs_y}{s_x}$, where r is Pearsons Coefficient and s_y and s_x are standard deviations of y and x.

First, we plot points of the data sets. Then illustrate by **least squares regression line** or trend line to help us deduce linear relationship between the bivariate data. Or we use the table as provided in the example below.



Example Given the data set, use least squares regression to derive the equation.

X	2	3	4	5	6
y	4	6	4	3	5

Solution

$y = a + bx$ where $b = rs_y/s_x$ and $a = \bar{y} - b\bar{x}$

Step 1. Complete the table.

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
2	4	-2	4	-0.4	0.16	0.8
3	6	-1	1	1.6	2.56	-1.6
4	4	0	0	-0.4	0.16	0
5	3	1	1	-1.4	1.96	-1.4
6	5	2	4	0.6	0.36	1.2
	Totals	**0**	**10**	**- 0.6**	**5.6**	**-1**
$\bar{x}=4$	$\bar{y}=4.4$					

Step 2. Calculate standard deviations of x and y.

$$s_x = \sqrt{\frac{10}{5-1}} = \sqrt{\frac{10}{4}} = \sqrt{2.5} = 1.581139... = 1.6$$

$$s_y = \sqrt{\frac{5.6}{5-1}} = \sqrt{\frac{5.6}{4}} = \sqrt{1.4} = 1.183216... = 1.2$$

Step 3. Solve for r when $r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$

$$r = \frac{1}{5-1} x \frac{-1}{1.6 \text{ x } 1.2}$$

$$= \frac{1}{4} x \frac{-1}{1.92} = \frac{-1}{7.68}$$

$$= 0.1302...$$

$$. = 0.13$$

Step 3. Solve for b in $b = \frac{rs_y}{s_x}$

$$b = \frac{0.13 \text{ x } 1.6}{1.2} = \frac{0.208}{1.2} = 0.17\dot{3} = 0.2$$

Step 4 Solve for a in $a = \bar{y} - b\bar{x}$.

a = 4.4 – 0.17 x 5= 4.4 – 0.85 = 3.6 (3.55 correct to 1dp)

Step 5 Substitute into y = a + bx for a and b.

y = 3.6 + 0.2x

Learning Activity 11.3.1.5 and 11.3.1.7

20 minutes

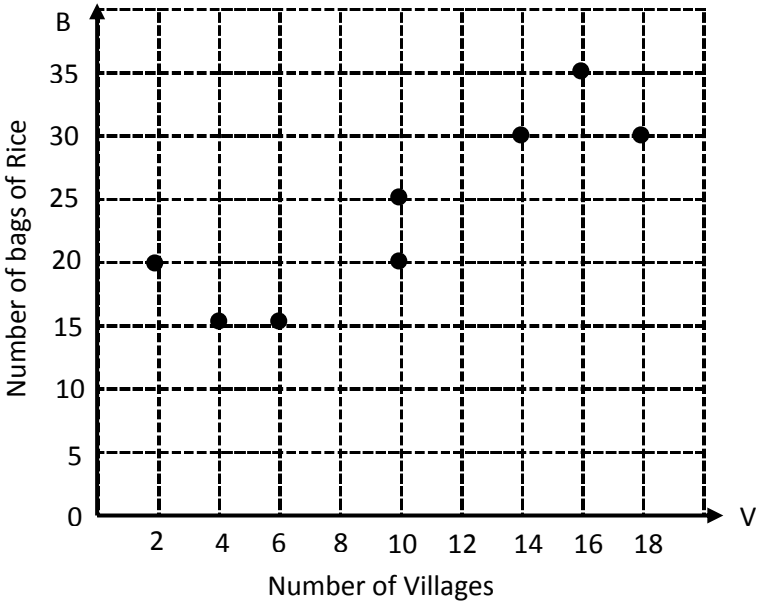
1. Below is the frequency distribution of the IQ of 40 students.

   a. Complete the table below by finding the cumulative frequency less than and the relative frequency of each class.

Classes	*f*
57 – 59	1
54 – 56	2
51 -53	4
48 – 50	6
45 – 47	10
42 – 44	17
39 – 41	3
36 – 38	4
33 – 35	2
30 – 32	1

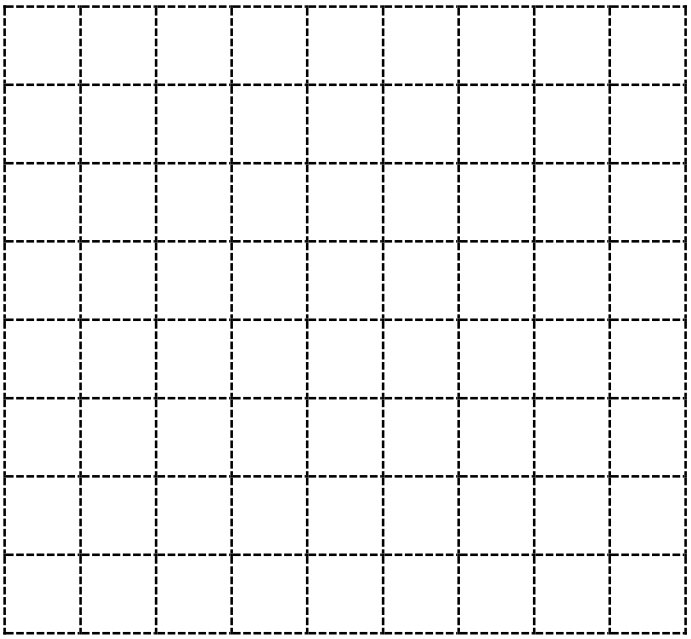
   b. Construct a cumulative frequency less than or OGIVE.

2. The graph shows the number of villages and the number of bags of 25kg rice distributed to each village during drought period in 2015. Derive equation of the scatter plot.



3. Draw a scatter plot of the data and state it's correlation. Then use least square regression to derive it's equation of the line – of – best - fit.

English	10	8	10	6	5	8	4	7
Maths	4	8	8	9	6	7	6	5



Summative Tasks 11.3.1

60 minutes

A. MULTIPLE CHOICE. Write the letter of your choice on the blank before each number.

_____ 1. What term refers to the body of language that deals with the collection, organization, presentation, and analysis of data?

a. statistics
b. probability
c. population
d. algebra

_____ 2. Which is a discrete data / variable?

a. weight of a boy
b. number of typewriter
c. volume of a container
d. average speed of a car

_____ 3. Which method of gathering data is enforced by law?

a. direct method
b. questionnaire method
c. experimental method
d. registration method

_____4. Which method of gathering data is costly and time consuming?

a. Interview method
b. Questionnaire method
c. Registration method
d. Experimental method

_____5. Which of the following are facts or information or observations under study?

a. sample
b. population
c. data
d. variables

_____6. What is a small portion or part of a population?

a. data
b. sample
c. parameter
d. statistic

_____ 7. What type of data is arranged into classes?

a. grouped data
b. ungrouped data
c. qualitative data
d. quantitative data

_____ 8. If the data are arranged from lowest to highest but not in tabular form, then the data are

a. grouped data
b. ungrouped data
c. qualitative data
d. quantitative data

_____9. Which of the following is commonly known as frequency polygon?

a. bar graph
b. pictograph
c. circle graph
d. line graph

*** For item numbers 10 to 20, refer to the given table.**

Class Interval	Frequency
46 – 50	3
41 – 45	1
36 – 40	2
31 – 35	15
26 – 30	6
21 – 25	12
16 – 20	8
11 – 15	3

Find:

_____10. The class width (*i*)

a. 5
b. 6
c. 8
d. 50

_____11. The number of observations (N)

a. 45
b. 50
c. 8
d. 46

_____12. The class which has 30% of the total observations

a. 46 – 50
b. 11 – 15
c. 31 – 35
d. 21 – 25

_____13. The class mark of the third class interval

a. 21
b. 25
c. 23
d. 25.5

_____14. The upper limit of the last class interval

a. 11.5
b. 15.5
c. 11
d. 15

_____15. The class with the highest frequency

a. 31 – 35
b. 46 – 50
c. 11 – 15
d. 41 – 45

_____16. The class boundaries of the first class interval

a. 46.5 – 50.5
b. 45.5 – 60.5
c. 46.5 – 60.5
d. 45.5 – 50.5

_____17. The number of classes

a. 6
b. 8
c. 12
d. 15

_____18. The class with the least frequency

a. 11 – 15
b. 36 – 40
c. 41 – 45
d. 46 – 50

_____19. The lower limit of the last class

a. 10.5
b. 11.0
c. 11.5
d. 10.0

_____20. The cumulative frequency of the class with the highest frequency

a. 29
b. 44
c. 46
d. 47

B. 1. The following marks are obtained by 50 students during a test.

(a) Complete the table based on the students marks.

5 4 6 5 4 6 6 5 4 7 6 5 6 4 5 5 6 6 8 5 4 2 8 6 8
3 6 4 7 7 3 4 7 4 1 7 4 5 6 4 5 8 7 6 9 7 5 4 6 2

Test Score	Tally	f	fX
		∑f=	∑fX=

(b) Add cumulative frequency column and draw cumulative frequency histogram of the data.

2. Construct a frequency table, a bar graph and a frequency polygon on the IQ test results of freshmen college students in a certain University in their entrance exam.

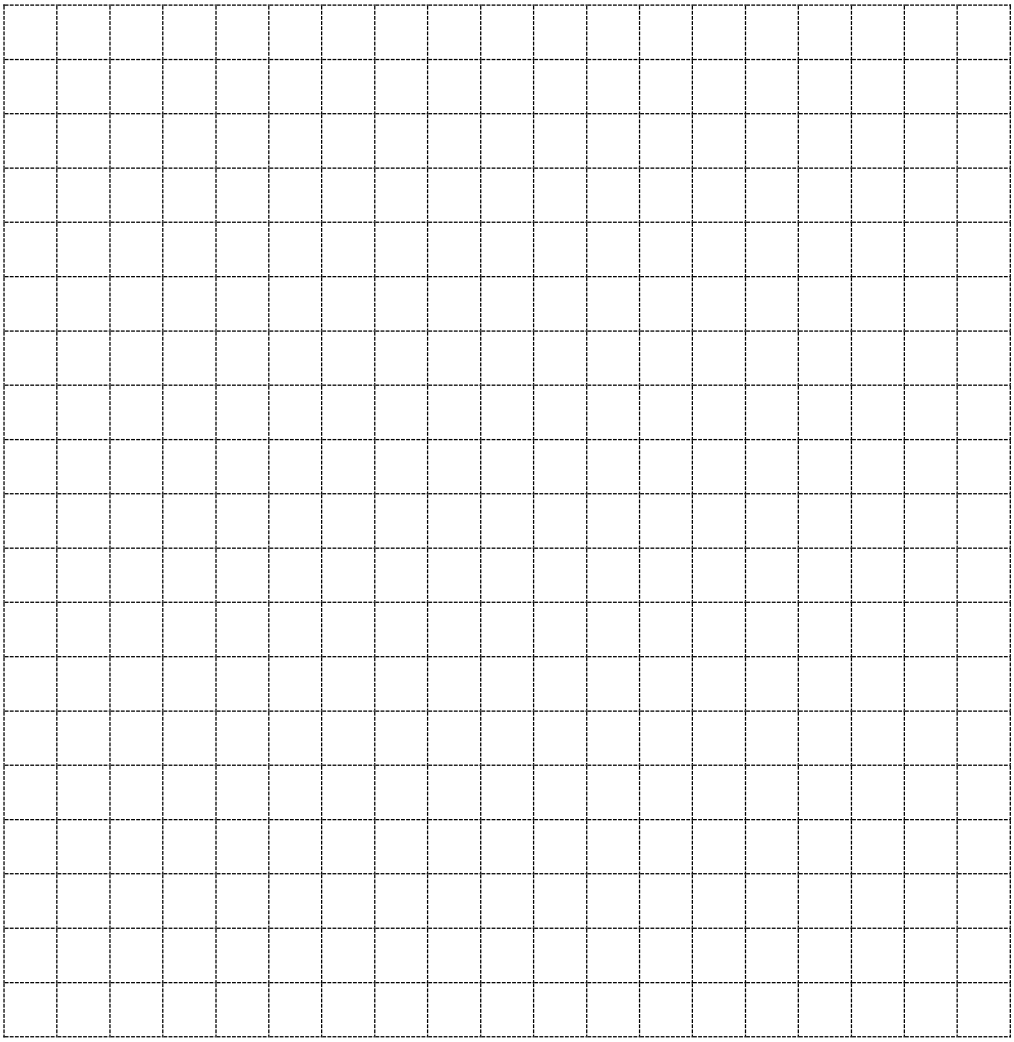
Complete the table with the following columns:
Class Interval, Frequency, Class Mark, CF< and Relative Frequency.(Use a separate sheet if necessary).

Note: Divide the range by 10 and start the lowest limit with a number that is divisible by the class size (*i*).

68	94	86	91	85	96	73	91
71	90	95	89	89	89	77	87
82	85	91	86	87	95	76	77
83	73	81	95	98	87	92	81
71	90	85	96	82	97	94	90

3. Plot a scattergram for the data and find the equation of the line – of - best – fit.

X	2	3	3	5	6	8	8	9	10	12	12	14	15	16
y	8	10	12	14	14	15	16	18	19	24	22	26	28	30



4. Apply least square regression technique to derive equation for the following data set.

X	2	4	6	8	10	12
y	35	28	20	14	8	2

Hint: Y = a + bx where $r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$ and $b = \frac{rs_y}{s_x}$, $a = \bar{y} - b\bar{x}$.

11.3. 2: MEASURES OF CENTRAL TENDENCY

Now that you have an idea on how to gather, organize and present data, let us explore ways of summarizing numerical data with a single number called the **average**.

A **measure of central tendency** is a single, central value that summarizes a set of numerical data called **average.** It is used to describe what is 'typical' in a set of data.

The central tendency or measure of centrality is the statistic that indicates an average value of a distribution. **Statistic** is any computed value obtained from a sample.
Several types of averages can be defined and the most commonly used type in statistics are the Mean, Median and Mode.

11.3.2.1 The Mean

The arithmetic mean can be used in almost all situations. It is particularly useful when comparing different populations, or estimating population values from sample values.

MEAN is the most common type of arithmetic average. The mean of the set of data is the sum of all the measurements divided by the number of measurements contained in the set of data. The symbol used to represent the mean average is $\overline{X}$.

One of the important characteristics of the mean is that it is easily affected by **outliers or extreme values**. An outlier can be very high or very low extreme value. It may also produce value which may not occur in practice; it may produce a value which does not reflect the distribution.

Advantages are it is easily understood. It is easily calculated in all forms numerical data. It takes account of all the data in the sample and is easy to manipulate algebraically.

Remember that there is only one mean for the set of data. We use the formula below in finding the mean of ungrouped data.

Mean Formula for Ungrouped Data

$$Mean(\overline{X}) = \frac{\Sigma X}{N}$$

Where ΣX is the sum of all data
N is the number of data

Example 1 The grades in Statistics of 10 students are shown in the table.

Mark	Roger	Kim	Rose	Ann	Liza	Dave	Myra	June	John
87	84	85	82	86	90	78	83	80	79

What is the mean average grade of the 10 students in their Statistics class?

$$Mean(\overline{X}) = \frac{\Sigma X}{N} = \frac{87+84+85+82+86+90+78+83+80+79}{10}$$

$\overline{X} = \frac{834}{10}$ $\overline{X} = 83.4$ The mean average of 10 students in their Statistics class is 83.4.

Example 2 The grades of Mary in her 5 exams are 85, 78, 82, 80 and 84. What must be her grade in the 6th quiz so that she will have an average of 83?

Solution

Using the formula for the Mean average

$$Mean(\overline{X}) = \frac{\Sigma X}{N} = \frac{85+78+82+80+84+n}{6} = 83$$

$$\frac{409+n}{6} = 83$$

$$83\ (6) = 409 + n$$

$$498 = 409 + n$$

$$498 - 409 = n$$

n = 89 Anne needs to obtain a grade of 89 in her 6th exam.

Example 3 Find the mean average of each group of data.

a. Group A: 2, 4, 6, 8, 10, 12, 14, 16, 18

b. Group B: 2, 4, 6, 8, 10, 12, 14, 16, 98

c. Compare the mean average of each group and explain.

Solution

a. $Mean(\overline{X}) = \frac{\Sigma X}{N} = \frac{2+4+6+8+10+12+14+16+18}{9}$ $\boxed{\overline{X} = 10}$

b. $Mean(\overline{X}) = \frac{\Sigma X}{N} = \frac{2+4+6+8+10+12+14+16+98}{9}$ $\boxed{\overline{X} = 18.89}$

c. Answers may vary. One may answer like this statement.

The mean average of group B is higher by 8.89 since the given data are almost identical except for the last values. The last data in group B which is 98 is an outlier (or very high extreme value) that greatly affects the mean average.

When the data values are grouped in a frequency distribution, the mean average can be computed using the **Mean Formula for Grouped Data**.

$$\overline{X} = \frac{\Sigma fX}{N}$$

Where ***f*** is the frequency of the class interval
X is the midpoint or class centre
N is the total number of observations

Example 4 The frequency distribution below shows the age of employees in ABC Corporation.

Classes (Age)	Frequency (*f*)	Class Mark (*X*)	*fX*
61 – 65	3	63	189
56 – 60	4	58	232
51 – 55	5	53	265
46 – 50	12	48	576
41 – 45	23	43	989
36 – 40	20	38	760
31 – 35	18	33	594
26 – 30	9	28	252
21 – 25	4	23	92
***i* = 5**	Σf = 98		ΣfX = 3949

a. Compute the mean average using the given data.

b. Which age group among the employees has the most number of frequency?

c. Which age group has the least number of employees?

d. In your own opinion, how will you interpret the age of the employees in ABC Corp.?

Solution

$$\bar{X} = \frac{\Sigma fX}{N} \qquad \bar{X} = \frac{3949}{98} \qquad \boxed{\bar{X} = \mathbf{40.30}}$$

a. The mean average age of 98 employees in ABC Corporation is **40.30.**

b. The age of employees with the highest frequency ranges from 41 to 45 years.

c. The age of employees with the lowest frequency ranges from 61 to 65 years.

d. Answers may vary. It may be concluded that 43. 88% of the age of the employees ranges from 36 to 45 years.

11.3.2.2 The Median

Revisiting example discussed in the mean, an outlier with a very high extreme value greatly affects the computed mean average. When there is an outlier in a given set of data and we aim to find the average, it is more appropriate to use the MEDIAN.

The **median** is the middlemost value in a set of data arranged in ascending or descending order. The median is another type of average and most appropriate to use when the middle value is desired. The symbol used to represent the median is $\tilde{X}$.

Just like the mean, there can only be one median in a set of data but unlike the mean, it is not influenced by extreme values. The median lies between the highest and lowest measurement where half of the data scores are located above the median and the other half is found below it when arranged in either ascending or descending order. And it can be determined in all situations.

Disadvantage of median is that it does not use all the values in the sample and it cannot be manipulated algebraically.

Median for Ungrouped Data

Example 1 The Attendance Monitoring System (AMS) shows the daily attendance of 35 students last week in their Mathematics class.

Monday	Tuesday	Wednesday	Thursday	Friday
34	30	28	31	27

Find the median of the given set of data.

Solution

To find the median, arrange the data in ascending order.
27, 28, 30, 31, 34

You may notice from the arranged data that the middle value is 30. We may therefore conclude that the median is 30. The median average of the students' attendance last week is 30.

Example 2 Janine's scores in 8 quizzes during the first quarter are 7, 8, 6, 9, 10, 8, 5, and 7. Find the median.

Solution

Arrange the data in increasing order: 5, 6, 7, 7, 8, 8, 9, 10
Since the given data is even, the median is the average of the two middle scores.

$$Median(\tilde{X}) = \frac{7+8}{2} = 7.5$$

Hence, the median of the set of scores is 7.5

Median for Grouped Data

In computing for the median of grouped data, the following formula is used

$$\mathbf{Median(\tilde{X}) = Lb_{\tilde{x}} + \left[\frac{\frac{\Sigma f}{2} - Cf <}{f_{\tilde{x}}}\right](i)}$$

Where, $\frac{\Sigma f}{2}$ is the median class found at $cf <$

$Lb_{\tilde{X}}$ is the lower class boundary of the median class

$f_{\tilde{X}}$ is the frequency of the median class

$cf <$ is the cumulative frequency of the class below the median class

i is the class interval or class size

Example 3 Calculate the median of the given grouped data

Pledges for the Victims of Earthquake in Nepal

Pledges in $	Frequency
9,000 – 9,999	3
8,000 – 8,999	6
7,000 – 7,999	8
6,000 – 6,999	13
5,000 – 5,999	21
4,000 – 4,999	25
3,000 – 3,999	18
2,000 – 2,999	32
1,000 – 1,999	30
0 – 999	14

Solution

Step 1: To calculate the median, complete the table by adding the cumulative frequency column.

Pledges for the Victims of Earthquake in Nepal

Pledges in $	Frequency	Cf<	
9,000 – 9,999	3	170	
8,000 – 8,999	6	167	
7,000 – 7,999	8	161	
6,000 – 6,999	13	153	
5,000 – 5,999	21	140	
4,000 – 4,999	25	119	
3,000 – 3,999	**18**	**94**	**Median Class**
2,000 – 2,999	32	76	
1,000 – 1,999	30	44	
0 – 999	14	14	
	$\sum f = 170$		

Step 2: Locate the median class

$\frac{\Sigma f}{2}$ is the median class found at $cf <$

$\frac{\Sigma f}{2} = \frac{170}{2} = 85$ Locate 85 at CF<

The median class is located at 3,000 – 3,999 since 85 is located in CF< 94 for the cumulative frequency of that class ranges from 77 to 94.

Step 3: From the median class, we identify the values of the following

$\frac{\Sigma f}{2} = 85$

$Lb_{\tilde{X}} = 2{,}999.5$

$f_{\tilde{X}} = 18$

$cf <= 76$ the cumulative frequency below (or less than) the median class

$i = 1000$

Step 4: Compute the median using the formula for grouped data

$$\text{Median}(\tilde{X}) = Lb_{\tilde{X}} + \left[\frac{\frac{\Sigma f}{2} - Cf <}{f_{\tilde{X}}}\right](i)$$

$$\tilde{X} = 2{,}999.5 + \frac{85 - 76}{18}(1000) -$$

$$\tilde{X} = 2{,}999.5 + 0.5(1000)$$

$$\tilde{X} = 2{,}999.5 + 500$$

$\tilde{X} = 3{,}499.5$ The median amount of pledge is \$3,499.50

Step 5: Check

The median of 3,499.50 falls within the class boundaries of 3,000 – 3,999 which is 2,999.5 – 3,999.5

11.3.2.3 The Mode

The **mode** is the type of average used during elections. It is also used in doing a survey about the most saleable product or even in doing feasibility studies.

It is the easiest type of average to determine and it can actually be found by inspection, and is not affected by outliers. The symbol used to represent the mode is $\hat{X}$.

The **MODE ($\hat{X}$)** is the value or score which occurs most frequently in the set of data.

Unlike the mean and median, a distribution can have one or more modes. Sometimes the distribution may not have any mode at all. The mode is the value with the greatest frequency.

If the distribution has two modes then it is a **bimodal** distribution. If the distribution has more than two modes, then it is a **multimodal** distribution.

However, mode is difficult to calculate for continuous frequency distributions; it may depend on the class intervals chosen for frequency distribution; and it does not use all the values in the sample data.

MODE for Ungrouped Data

Steps in finding the Mode for Ungrouped Data.

1) Select the score (data) that appears most often in the set of data.

2) If there appears two or more score/data with the same number of times, then each of these values is a mode.

3) if every score(data) appears the same number of times, then the data has no mode.

Example 1 Find the mode of the given sets of data.

a. Set C: 8, 6, 9, 3, 5, 8, 1, 7, 8

Solution

Distribution 1, 3, 5, 6, 7, 8, 8, 8,9
The mode is 8 and called unimodal since there is only 1 mode.

b. Set D: 14, 15, 10, 14, 17, 10, 19

Solution

Distribution 10, 10, 14, 14, 15, 17, 19
The modes are 10 and 14. Thus, it is bimodal.

c. Set E: 1, 8, 9, 2, 6, 3, 10, 5

Solution

There is no mode in the given data since all data score appear only once.

d. Set F: 3, 4, 3, 7, 3, 8, 9, 6, 7, 7, 4, 4, 6, 6, 2, 5, 8, 5, 8

Solution

There are fives modes in this set of data (3, 4, 6, 7, and 8). This is called multi-modal.

MODE for Grouped Data

The mode for grouped data can be computed using this formula

$$\text{Mode}(\hat{X}) = Lb_{mo} + \frac{D_1}{D_1 + D_2}(i)$$

Where Lb_{mo} is the lower class boundary of the modal class

D_1 is the difference between the frequencies of the modal class and the upper class with higher class

D_2 is the difference between the frequencies of the modal class and the upper class with lower class

i is the class size or class interval

Example 2 Compute the mode using the frequency table in page 23.

The frequency distribution below shows the age of employees in ABC Corporation.

Classes (Age) x	Frequency (f)	Class Mark (X)	fX	
61 – 65	3	63	189	
56 – 60	4	58	232	
51 – 55	5	53	265	
46 – 50	12	48	576	
41 – 45	**23**	**43**	**989**	**Modal Class**
6 – 40	20	38	760	
31 – 35	18	33	594	
26 – 30	9	28	252	
21 – 25	4	23	92	
$i = 5$	$\Sigma f = 98$		$\Sigma fX = 3949$	

Solution

The modal class is the class with the highest frequency. Thus, the modal class is 41 – 45.

Lb_{mo} = 40.5 i = 5

D_1 = *fmo – f(class higher class)* $D_1 = 23 - 12$ $D_1 = 11$

D_2 = *fmo – f(class lower class)* $D_2 = 23 - 20$ $D_2 = 3$

$$\text{Mode}(\hat{X}) = \text{Lb}_{mo} + \frac{D_1}{D_1 + D_2}(i)$$

$$\text{Mode}(\hat{X}) = 40.5 + \frac{11}{11+3}(5) \qquad \text{Mode}(\hat{X}) = 40.5 + 3.93$$

$$\text{Mode}(\hat{X}) = 44.43$$

Thus, the mode of 44.43 falls within the class boundaries of 41 – 45 which is 40.5 – 45.5.

Learning Activity 11.3.2.1 – 11.3.2.3

 60 minutes

1. Compute the mean, median and mode of the following sets of data.

   a. Set X: 43, 32, 26, 38, 35, 43, 45, 42, 43

   b. Set Y: 25, 18, 22, 26, 24, 22, 19, 25, 27

   c. Set C: 87, 82, 86, 78, 94, 88, 85

2. Below are the scores of students in their final exams.

Score	*f*	*X*	*fX*	*Cf<*
	3			
55 – 58	4			
51 – 54	6			
47 – 50	8			
43 – 46	11			
39 – 42	10			
35 – 38	8			
31 – 34	5			
27 – 30	2			
	Σf =		ΣfX = 3949	

   a. Complete the table by filling in the possible values per column.

b. Find the mean, median and mode of the set of data.

c. Compare the mean, median and mode obtained from the set of data.

d. Which of the three measures of central tendency represents the average of the set of data? Why?

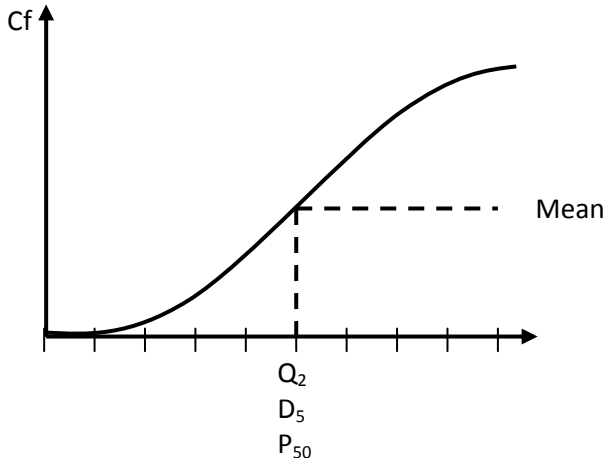
11.3.2.4 The Percentile and Quartile (Quantiles)

You have just learned about measures of centrality or averages. In this lesson, you will learn about measures of position. Did you try to compare your academic performance with that of your classmate or your friend?

In measures of central tendency, you have gained knowledge and deeper understanding about the characteristics of a data set. However, to have more knowledge about the data set, we may divide it into more parts of equal sizes.

If we have a large amount of data, we will focus on dividing it into quantiles; that is often into a hundred parts (percentile), ten parts (decile) and four parts (quartile).

Cumulative Frequency Curve



The **Percentiles** are the ninety-nine score points which divide the distribution into one hundred equal parts. It characterizes the values according to the percentage below the given score.

Percentile for Ungrouped Data

The percentiles determine the value for 1%, 2%, ..., up to 99% of the data set. This means that the first percentile (denoted by $P_{1)}$ separates the lowest 1% from the other 99% and so on.

P_{25} or 25^{th} percentile indicates that 25% of the data have values less than or equal to $P_{25.}$ P_{50} means that 50% of the data have values less than or equal to 50%.

Percentile

$$P_k = \frac{k(n+1)}{100}$$

Example 1 Find the 40^{th} percentile or P_{40} of the following raw data.
24, 31, 18, 63, 45, 28, 29, 58, 26, 30

Solution

Arrange the data from lowest to highest.
18, 24, 26, 28, 29, 30, 31, 45, 58, 63

To find its P_{40} position, we use the formula $P_k = \frac{k(n+1)}{100}$, then round off to the nearest whole number.

We use $P_{40} = \frac{40(10+1)}{100}$

$$P_{40} = \frac{40(11)}{100}$$
$$P_{40} = 4.4 \approx 4$$

This means that P_{40} is the 4^{th} data score in the distribution
Therefore, $P_{40} = 28$

We can now conclude that 40% of the data set have values less than or equal to 28.
18, 24, 26, 28, 29, 30, 31, 45, 58, 63

Example 2 The following sets of data are arranged in ascending order.

5	12	13	16	17	18	18	21	23	26	28	29	32	34	37
39	40	41	42	44	46	48	51	52	55	56	57	59	62	65

Find:

a. P_{25} b. P_{62} c. P_{85}

Solution

There are 20 scores in the given data set. So N = 30 scores

a.

$$P_{25} = \frac{25(30+1)}{100}$$
$$= \frac{25(31)}{100}$$
$$= \frac{775}{100}$$
$$= 7.75$$
$$\approx 8$$

P_{25} is the 5th data score which is 26.

b.

$$P_{62} = \frac{62(30+1)}{100}$$
$$= \frac{62(31)}{100}$$
$$= \frac{1922}{100}$$
$$= 19.22$$
$$\approx 19$$

P_{62} is the 13th score which is 42.

c.

$$P_{85} = \frac{85(30+1)}{100}$$
$$= \frac{85(31)}{100}$$
$$= \frac{2635}{100}$$
$$= 26.35$$
$$\approx 26$$

P_{85} is the 26th data score which is 56.

Percentile for Grouped Data

The percentile of grouped data is used to characterize values according to the percentage below them. In finding the percentiles of grouped data, we revisit the method of finding the median on page 24 because it is similar to that of finding the median.

$$P_k = Lb + \left(\frac{\frac{kN}{100} - Cf_b}{f_{p_k}} \right)(i)$$

Where, *Lb* is the lower class boundary of the k^{th} percentile class
N is the total frequency
Cf_b is the cumulative frequency before or below the k^{th} percentile class
F_{Pk} is the frequency of the percentile class
K is the n^{th} percentile where n = 1, 2, 3, ..., 98, and 99
i size of the class interval

Example Calculate the (a) 27^{th} percentile and (b) 70^{th} percentile of the scores of 50 students in their Mathematics test scores.

Scores	Frequency
81 – 90	4
71 – 80	7
61 – 70	12
51 – 60	8
41 – 50	10
31 – 40	6
21 – 30	3

Solution

To calculate P_{27} and P_{70}, we first complete the table by adding two more columns for Lower class boundaries (L_b) and cumulative frequency less than ($Cf<$).

Scores	Frequency	Lower Boundaries (L_b)	Cumulative Frequency < ($C_f<$)	
81 – 90	4	80.5	50	
71 – 80	7	70.5	46	
61 – 70	12	60.5	39	⇦ P_{70} class
51 – 60	8	50.5	27	
41 – 50	**10**	**40.5**	**19**	⇦ P_{27} class
31 – 40	6	30.5	9	
21 – 30	3	20.5	3	
	N = 50			

(a) To locate P_{27} class just like in locating the median class we find $\frac{kN}{100}$.

P_{27} class = $\frac{27(50)}{100}$

P_{27} class = **13.5**

We locate this value to $Cf<$ where 13.5^{th} score is contained.
The 10^{th} to 19^{th} scores are found on 41 – 50 class interval.
The P_{27} = 13.5 belong to the same class interval of 41 – 50.

We then use the grouped data formula:

$$P_{27} = Lb + \left(\frac{\frac{kN}{100} - Cf_b}{f_{P_k}}\right)(i)$$

$Lb = 40.5$ $\quad Cf_b = 9$

$k = 27$ $\quad f_{P_k} = 10$

$N = 50$ $\quad i = 10$

$$= 40.5 + \left(\frac{\frac{27(50)}{100} - 9}{10}\right) x10$$

$$= 40.4 + \left(\frac{13.5}{10}\right)10$$

$P_{27} = 54$

$$= 40.5 + 13.5$$

$$= 54$$

From the computed value of $P_{27} = 44$, we then conclude that 27% of the students got a score less than or equal to 41 – 50.

(a) To locate P_{70} we again use $\frac{kN}{100}$.

P_{70} class = $\frac{70(50)}{100}$

P_{70} class = **35**

We locate this value to *Cf*< where 35th score is contained.
The 28th to 39th scores are found on 61 – 70 class interval.
The $P_{70} = 35$ belong to the same class interval of 61 – 70.

We then use the grouped data formula:

$$P_{70} = Lb + \left(\frac{\frac{kN}{100} - Cf_b}{f_{P_k}} \right)(i)$$

$Lb = 60.5$ $\quad Cf_b = 27$
$k = 70$ $\quad f_{P_k} = 12$
$N = 50$ $\quad i = 10$

$$= 60.5 + \left(\frac{\frac{70(50)}{100}}{12} \right)10$$

$$= 60.5 + \frac{8}{12}(10)$$

$$= 60.5 + 6.7$$

$$= 67.7$$

$P_{70} = 67.7$

From the computed value of $P_{70} = 67.17$, we then conclude that 70% of the students got a score less than or equal to 61 – 70.

In percentile, we divide the data score by 100. We can also divide the distribution by 4 and we call it QUARTILE.

The **Quartiles** are the score points which divide the distribution into four equal parts. 25% of the distribution is below the first quartile, 50% are below the second quartile, and 75% are below the third quartile. Q_1, Q_2, and Q_3 are called the 1st, 2nd, and 3rd quartiles respectively. Q_1 is the lower quartile while Q_3 is the upper quartile and Q_2 is also the median.

Quartile for Ungrouped Data

Note that $Q_1 < Q_2 < Q_3$

a. 25% of the data has a value ≤ Q_1.

b. 50% of the data has a value ≤ Q_2 or $\overline{X}$

c. 75% of the data has a value ≤ Q_3.

Example 1 Find the 1st quartile, 2nd quartile and 3rd quartile of the following raw data.
24, 31, 18, 63, 45, 26, 29, 58, 26, 30, 35

Solution

If there are odd numbered given data, we can easily locate the values of Q_1, Q_2 and Q_3.

i. Arrange the data from lowest to highest.
18, 24, 26, 26, 29, 30, 31, 35, 45, 58, 63

ii. Since the least value in the given data is 24 and the greatest value is 63, the middle value is 30. Therefore Q_2 or $\overline{X}$ = 30.

iii. The lower quartile Q_1 is between the middle value and the least value. So Q_1 = 26.

iv. The upper quartile Q_3 is between the middle value and the greatest value. So Q_3 = 45. Note that we do not get the average of the two middle scores between the middle value and the largest value. We select the data with the higher value instead.

We can also find the values of the three quartiles using the formula similar to percentile for ungrouped data.

v. To find its Q_1 position, we use the formula $Q_k = \frac{k(n+1)}{4}$, then round off to the nearest whole number.

We use $Q_1 = \frac{1(11+1)}{4}$

$Q_1 = 3$ This means that Q_1 is the 3rd data score.

Therefore, Q_1 = 26

vi. To find its Q_2 position, we have $Q_2 = \frac{2(11+1)}{4}$

$Q_2 = 6$

This means that Q_2 is the 6^{th} data score.

Therefore, $Q_2 = 30$

vii. To find its Q_3 position, we have $Q_3 = \frac{3(11+1)}{4}$

$Q_3 = 9$

This means that Q_3 is the 9^{th} data score.

Therefore, $Q_3 = 45$

You may notice that with or without the formula, we may obtain the same data scores in locating the three quartiles.

Quartile for Grouped Data

In finding the quartiles of grouped data, we use the same procedure in locating the percentile data. The formula that we use is

$$Q_k = Lb + \left(\frac{\frac{kN}{4} - Cf_b}{f_{Qk}} \right)(i)$$

Where, *Lb* is the lower class boundary of the k^{th} quartile class

N is the total frequency

Cf_b is the cumulative frequency before or below the k^{th} quartile class

F_{Pk} is the frequency of the quartile class

K is the n^{th} quartile where n = 1, 2, and 3

i size of the class interval

Example Calculate the first, second and third quartile of the scores of 50 students in their Mathematics test scores.

Scores	Frequency
81 – 90	4
71 – 80	7
61 – 70	12
51 – 60	8
41 – 50	10
31 – 40	6
21 – 30	3

Solution

To calculate the three quartiles , we first complete the table by adding two more columns for Lower class boundaries (L_b) and cumulative frequency less than ($Cf<$).

Scores	Frequency	Lower Boundaries (L_b)	Cumulative Frequency < ($Cf<$)	
81 – 90	4	80.5	50	
71 – 80	7	70.5	46	
61 – 70	**12**	**60.5**	**39**	⇦ Q_3 class
51 – 60	**8**	**50.5**	**27**	⇦ Q_2 class
41 – 50	**10**	**40.5**	**19**	⇦ Q_1 class
31 – 40	6	30.5	9	
21 – 30	3	20.5	3	

N = 50

(a) To locate Q_1 class we find $\frac{kN}{4}$.

Q_1 class = $\frac{1(50)}{4}$

Q_1 class = **12.5** We locate this value to *Cf*< where the 12.5th score is contained.
The 10th to 19th scores are found on 41 – 50 class interval.
The Q_1 class = 12.5 belong to the same class interval of 41 – 50.

We then use the grouped data formula:

$$Q_1 = Lb + \left(\frac{\frac{kN}{4} - Cf_b}{f_{Qk}} \right)(i)$$

$Lb = 40.5$ $Cf_b = 9$
$k = 1$ $f_{P_k} = 10$
$N = 50$ $i = 10$

$$Q_1 = 40.5 + \frac{\frac{1(50)}{4} - 9}{10}(10) = 40.5 + \frac{12.5 - 9}{10} \text{ x } 10 = 40.5 + \frac{3.5}{10} \text{ x } 10$$

$$Q_1 = 40.5 + 3.5$$

$Q_1 = 44$

From the computed value of $Q_1 = 44$ and since $Q_1 = P25$, we then conclude that 25% of the students got a score less than or equal to 41 – 50.

(b) To locate Q_2 class we find $\frac{kN}{4}$.

Q_2 class = $\frac{2(50)}{4}$

Q_2 class = **25** We locate this value to *Cf*< where the 25th score is contained.
The 20th to 27th scores are found on 51 – 60 class interval.
The Q_2 class = 25 belong to the same class interval of 51 – 60.

We then use the grouped data formula:

$$Q_2 = Lb + \left(\frac{\frac{kN}{4} - Cf_b}{f_{Q_k}} \right)(i)$$

$Lb = 50.5$ $Cf_b = 19$
$k = 2$ $f_{Q_k} = 8$
$N = 50$ $i = 10$

$$= 50.5 + \left(\frac{\frac{2(50)}{4} - 19}{8} \right)(10)$$

$$= 50.5 + \frac{6}{8}(10)$$

$$= 50.5 + 7.5$$

$$= 58$$

$Q_1 = 58$

From the computed value of $Q_2 = 58$ and since $Q_2 = P_{50}$, we then conclude that 50% of the students got a score less than or equal to 51 – 60.

(b) To locate Q_3 class we find $\frac{kN}{4}$.

Q_3 class = $\frac{3(50)}{4}$

Q_3 class = **37.5**

We locate this value to *Cf*< where the 37.5th score is contained.
The 28th to 39th scores are found on 61 – 70 class interval.
The Q_3 class = 37.5 belong to the same class interval of 61 – 70.

We then use the grouped data formula:

$$Q_3 = Lb + \left(\frac{\frac{kN}{4} - Cf_b}{f_{Q_3}}\right)(i)$$

$Lb = 60.5$ $\quad Cf_b = 27$

$k = 3$ $\quad f_{Q_k} = 12$

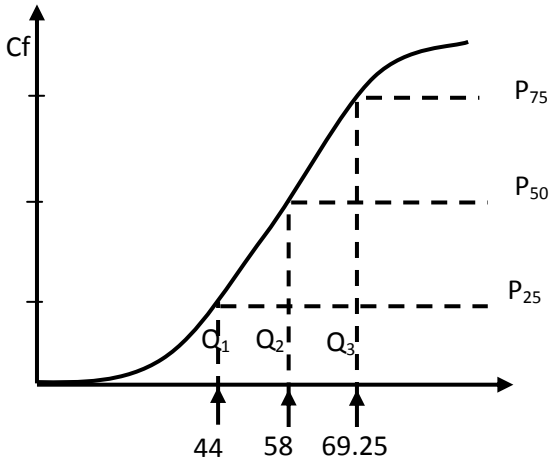
$N = 50$ $\quad i = 10$

$$Q_3 = 60.5 + \left(\frac{\frac{3(50)}{4} - 27}{12}\right)(10) = 60.5 + \left(\frac{\frac{150}{4} - 27}{12}\right)(10) = 60.5 + \left(\frac{37.5 - 27}{12}\right)(10)$$

$$Q_3 = 60.5 + \left(\frac{10.5}{12}\right)(10) = 60.5 + \frac{105}{12} = 60.5 + 8.75$$

$Q_3 = 69.25$

From the computed value of $Q_3 = 69.35$ and since $Q_3 = P_{75}$, we then conclude that 75% of the students got a score less than or equal to 61 – 70.



Learning Activity 11.3.2.4

60 minutes

1. Answer the following completely.

The daily wages of eight workers in a garment factory are:

380, 410, 400, 420, 360, 465, 438, 524

a. P_{80}

b. P_{14}

c. Q_1

d. Q_3

2. The frequency table below shows the monthly salary of 40 employees in GRACE Company. Use the given information in the table to answer the questions that follow.

Monthly Salary of Employees in GRACE Company

Salary	Frequency
9,001 – 10,000	6
8,001 – 9,000	8
7,001 – 8,000	9
6, 001 – 7,000	12
5,001 – 6,000	3
4,001 – 5,000	2

Calculate the 60^{th} percentile and the 3^{rd} quartile of the monthly salary of the employees.

11.3.2.5 Relation of the Mean, Median and Mode

Revisiting the mean, median and mode as measures of central tendency, let us now identify the characteristics of each and discover their similarities and differences. These are all averages but they have different distinctions on how they are used.

We typically use the mean when we compute the average grade of a student or in judging the contestants in a beauty pageant, and the like. On the other hand, we use the median if we are after the middle score in a distribution while the mode is used during election or on survey when we wish to identify the most saleable product in the market.

To understand this further, try to compare the mean, median and mode of the following examples.

Example 1 Compute the mean, median and mode of the scores in Science of the 3 students in their assignments.

Arvin	:	91	79	81	84	78	87	88
Amiel	:	46	59	61	84	85	86	167
Daniel	:	79	84	89	84	84	76	92

Solution

Name of students	MEAN	MEDIAN	MODE
Arvin	84	84	No mode
Amiel	84	84	No mode
Daniel	84	84	84

Notice that the three students obtained the same mean average and a median of 84. If these will be encoded in their report cards, we may think and conclude that they have the same level of intelligence in their Science subject if our basis will only be the grades that appear in their record. Therefore we may say that our interpretation or conclusion is wrong.

What is the most appropriate type of average to use?

In cases like this, it is important that we analyze and identify the given data first before we give our decision. The following information may be considered as our guide in identifying the appropriate type of average.

The mean can be used if there is no outlier since the mean is greatly affected by extreme values. Likewise, the mean can also be used if there is no mode or there is multiple number of modes.

The median is most appropriate to use as a when there is an **outlier** and when the middle value is desired.

The mode is the easiest type of average to use because it can actually be found by inspection. It is appropriate to use if one or few data appear most often in a distribution. It is not advisable to use if there are very few data.

Example Consider the following data and identify which measure of central tendency is most appropriate to use.

1. 8, 9, 10, 12, 12, 14, 15, 17, 90

2. 49, 56, 58, 60, 62, 64, 65, 68, 70, 72, 73, 75, 78, 81, 83, 83, 86

3. 3, 5, 6, 7, 8, 9, 9

Solution

1. Since there is an outlier (a very high score of 90), the **Median** is most appropriate although the mode is 12. The mean average cannot be used in the presence of an outlier, which is 90. Likewise, the mode of 12 is not applicable because there are only very few data.
   Where mean $= \frac{8+9+10+12+12+14+15+17+90}{9} = \frac{187}{9} = 20.\dot{7}$, which does not reflect a middle value.

2. Since there is no outlier and the mode is close to the extreme score, the **mean** average is most appropriate.

$$\bar{x} = \frac{49+56+58+60+62+64+65+68+70+72+73+75+78+81+83+83+86}{17}$$
$$= \frac{1183}{17}$$
$$= 69.6$$
$$\approx 70$$

We round 69.6 to 70 because there is no decimal score in the set of data.

3. Although there is a mode of 9 and there is no outlier, the **mean** average is most appropriate to use in this set of data. Mode of 9 is not applicable for there are only very few data, and is an extreme.

Outlier is an extreme value that is outside other values in a set of data.

For **grouped data**, there is a possibility that there will be two or more modal classes. In this case, you will not compute for the two or more modes. The empirical relation of the mean median and mode is described by this formula

Mode = 3(Median) – 2(Mean)

Example Find the (a) mean, (b) median and (c) mode of the frequency distribution table below.

Height of Flexible Open and Distance Education Students in Papua New Guinea

Height (in cm)	Frequency
171 – 175	6
166 – 170	10
161 – 165	11
156 – 160	11
151 – 155	10
146 – 150	8

Solution

To find the three averages, we need to complete the table by adding **X**, ***f*X**, ***Lb***, and ***Cf<*** columns.

Height of Flexible Open and Distance Education (FODE) Students in Papua New Guinea

Height (in cm)	Frequency	*X*	*f X*	*Lb*	*Cf<*	
171 – 175	6	173	1038	170.5	56	
166 – 170	10	168	1680	165.5	50	
161 – 165	11	163	1793	160.5	40	
156 – 160	**11**	**158**	**1738**	**155.5**	**29**	**Median Class**
151 – 155	10	153	1530	150.5	18	
146 – 150	8	148	1184	145.5	8	
	N= $\sum f$ =56		ΣfX = 8963			

a. MEAN $(\overline{X}) = \frac{\Sigma fX}{N}$

MEAN $(\overline{X}) = \frac{8963}{56}$ **MEAN $(\overline{X})$ = 160.05 cm**

The mean average of the height of FODE students in Papua New Guinea is 160.05 cm.

b. MEDIAN

Locate the median class

$\frac{\Sigma f}{2}$ is the median class found at $cf <$

$\frac{\Sigma f}{2} = \frac{56}{2} = 28$ Locate 28 at CF<

The median class is located at 156 – 160 since 28 is located in CF< 29 for the cumulative frequency of that class ranges from 19 to 29.

From the median class, we identify the values of the following

$\frac{\Sigma f}{2} = 28$

$Lb_{\tilde{X}} = 155.5$

$f_{\tilde{X}} = 11$

$cf <= 18$ the cumulative frequency below (or less than) the median class

$i = 5$

Compute the median using the formula for grouped data.

$$\text{Median}(\tilde{X}) = \text{Lb}_{\tilde{X}} + \left[\frac{\frac{\Sigma f}{2} - \text{Cf} <}{f_{\tilde{X}}}\right](i)$$

$$\tilde{X} = 155.5 + \frac{28 - 18}{11}(5)$$

$$\tilde{X} = 155.5 + 0.91(5)$$

$$\tilde{X} = 155.5 + 4.55$$

$$\tilde{X} = \mathbf{160.05}$$

The median average height of FODE students in Papua New Guinea is 160.05 cm which is within the boundaries of 156 – 160 class that is 155.5 – 160.5.

c. MODE

Notice that the given data has two classes with the highest frequency. We therefore conclude that to use the original formula in solving the mode is not applicable. Instead, we use the formula regarding the empirical relation of the mean, median and mode found in page 39.

Mode = 3(Median) – 2(Mean)

The computed mean and median are: **MEAN** $(\overline{X})$ **= 160.05 cm, MEDIAN** $(\tilde{X})$ = **160.05**

$$\text{Mode}(\hat{X}) = 3\tilde{X} - 2\overline{X}$$
$$= 3(160.05) - 2(160..05)$$

$$Mode(\hat{X}) = 160.05$$

The mode average height of FODE students in Papua New Guinea is 160.05 cm which happen to have the same mean and median.

11.3.2.6 Normal and Skewed Distributions

Distribution is an array of scores of an observation in ascending or descending order. A normal distribution occurs when the **mean**, the **mode** and the **median** of the distribution have the same value or are **equal**.

Here the following symbols are used to represent means, $\bar{x}$ and μ. However, in further statistics $\bar{x}$ is used for sample mean and μ is used to indicate population mean.

The mean is found by the formula $\bar{x} = \frac{\sum x}{N}$ or $\bar{x} = \frac{\sum x}{n}$.

If the frequency distribution is tabulated, the following formulae can be used $\bar{x} = \frac{\sum fx}{N}$ or $\bar{x} = \frac{\sum fx}{n}$.

The formula for calculating standard deviation (sd) σ is

$$\sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{N}} \text{ or } \sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{n}}.$$

The mean and the standard deviation are parameters of the density curve, and the shape is symmetrical about the mean in normal density curve.

Normal Distribution

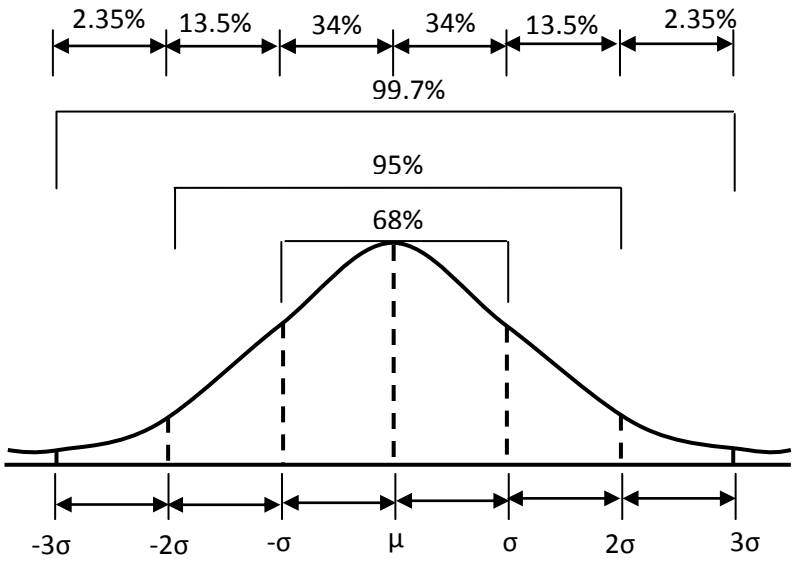
The normal distribution is best suited for data that, at the minimum, meets the following conditions:

- There is strong tendency for the data to take on central value.
- Positive and negative deviations from this central value are equally likely.
- The frequency of the deviations falls off rapidly as we move further away from the central value.

The ogive of the normal distribution (density) curve is symmetrical about the mean. In any normal distribution of an observation:

- 68% of the observations fall within σ of the mean μ
- 95% of the observations fall within 2σ of the mean μ
- 99.7% of the observations fall within 3σ of the mean μ

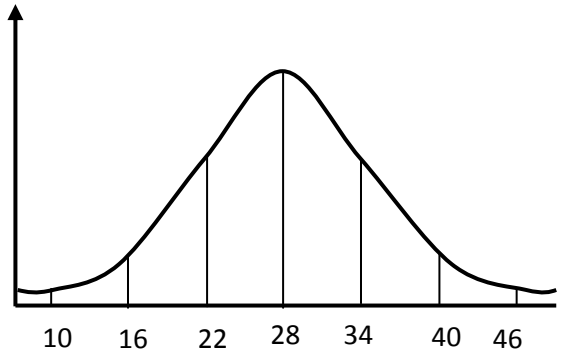
These findings were based on the studies of DeMoivre (1733) who derived the mathematical equation of the normal curve and Gauss (1777-1855) who derived the equation from a study of errors.



* μ stands for mean and σ stands for standard deviation (sd).

The normal distribution depends on mean and standard deviation (parameters); if the mean is affected by an outlier, use median value of the distribution to compute standard deviation. This is to ensure that 3σ above and below the median is contained in the data set. Using the mean may result in the 3σ above and below the mean to exceed the data set.

Suppose a school data in, say mathematics examination, has an assumed mean of 28 and a standard deviation of 6 in a normally distributed data. There is a total of 180 students.



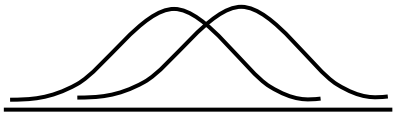
Since the data is normally distributed, the following results (and more) is likely:

- 122 students score will fall between 22 and 34 (68%)
- 171 students score will fall between 16 and 40 (95%)
- 86 students score will fall between 16 and 28 (47.5%)
- 4 students score will fall between 40 and 46 (2.35%)
- 151 students will pass if the cut-off mark is 22 (83.85%)

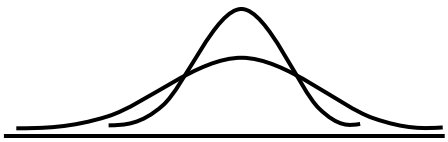
If in the following year the school performs well with scores normally distributed, the number of students may vary based on the examination class population. The number of passes also depends on the cut-off mark for that year.

Comparing two sets of data or observations

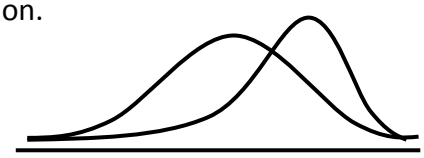
Two observations with same σ but different $\bar{x}$. Observation with smaller $\bar{x}$ is on the left.



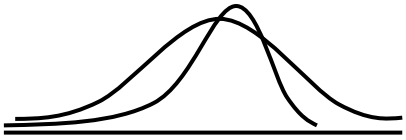
Two observations with same $\bar{x}$ but different σ. The observation with larger σ is wider.



Two observations with different $\bar{x}$ and different σ. One observation is a normal distribution, the other is a skewed distribution.



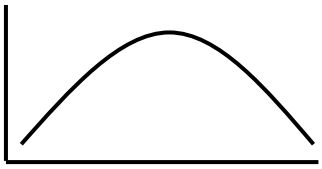
Two observations with the same $\bar{x}$. One is normal and the other is a negatively skewed distribution.



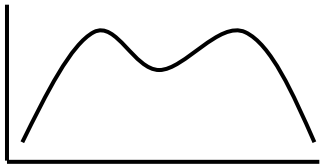
Note that skewness is determined by its direction of tail. A Positive Skew if the tail goes positive direction or to the right, a Negative Skew if the tail goes to the negative direction or to the left.

Distribution Types

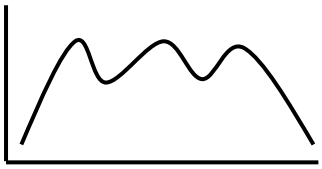
When a variable is measured across a large population, its frequency curve usually fits one of the following patterns.



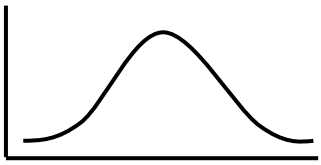
Unimodal distribution



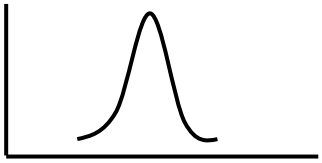
Bimodal distribution



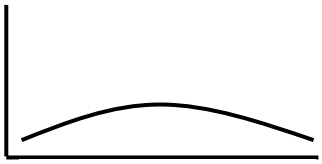
Multimodal distribution



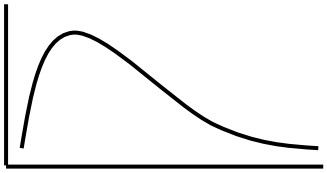
Normal distribution



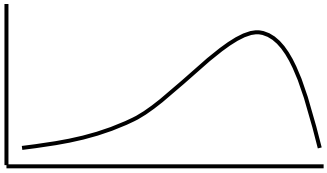
Small dispersion



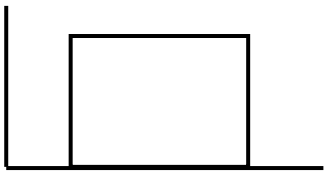
Large distribution



Positive or skewed right distribution



Negative or skewed left distribution



No mode

Analysis and Interpreting Data - The School Perspective

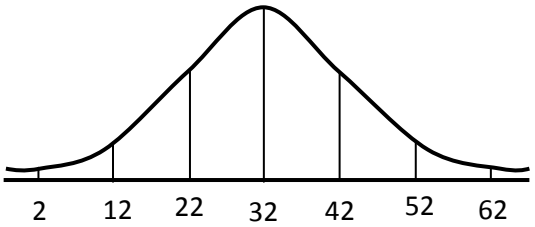
The approach is common when norm assessment method is concerned than Criterion assessment, z-score, stanine or IQ methods.

Bellow is a distribution of Science Examination results. The science examination is out of 50 marks. There were 45 students in grade 10A class.

11	13	16	18	19	19	21	21	22	23	24	24	25	26	26
27	28	29	29	32	32	33	34	34	35	36	37	37	37	38
38	38	39	40	41	42	42	43	43	45	46	47	48	48	50

The mean and standard deviation are given to the nearest whole number. We can use the symbols $\bar{x}$ and s for mean and standard deviation because a class data set is a sample of the school data set.

$\sum x = 1456$, $n = 45$, $\bar{x} = 32$, $\sigma_{n-1} = 10$



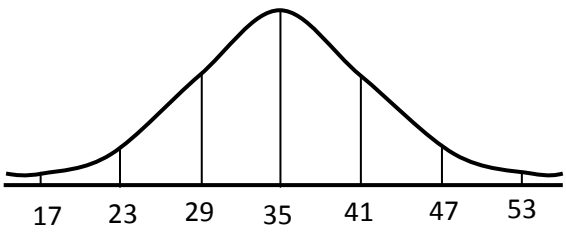
68% of students score between 22- 42 which is 31 (30.6). And 75% of students scored between 12 – 52 which is 34 students; and 99.7% scored between 2- 62 which is 45 (44.856) students.

But there are no raw scores greater than 50 and also the test total is 50. However the σ of 10 is so large, that is there is a big spread of results and the data is bimodal, therefore $\bar{x}$ + 2σ and $\bar{x}$ + 3σ give us scores greater than 50. With this observation, there won't be any score above 52 which is $\bar{x}$ + 2σ.

Bellow is a distribution of Science Examination results of 45 students of grade 10B class.

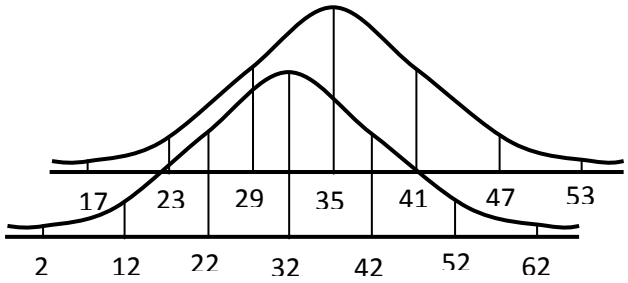
21	23	26	28	29	29	31	31	32	32	32	32	32	33	33
33	33	33	33	34	34	34	34	34	34	34	35	35	36	36
36	36	37	37	38	38	39	39	40	40	43	45	46	47	48

$\sum x = 1601$, $\bar{x} = 36$, $n = 45$, $\sigma_{n-1} = 6$



68% of students score is between 29- 41 which is 31 (30.6). And 75% of students scored between 23 – 47 which is 34 students; and 99.7% scored between 17- 53 which is 45 (44.856) students.

By inspection, if we have a couple more scores below 33, the mean would likely be 33. Then the mean, mode and median are likely to be equal so we will obtain a perfect normal distribution.



Comparing the two normal distributions, we can say that Class B had better sample mean $\bar{x}$, than Class A. If the national examination mean is 32, Class B is likely to obtain a higher **mean rating index** (MRI) based on higher number of Upper Pass (UP) + Credit (C) + Distinction (D) the class B is likely to get. The sum is then divided by the class population to obtain MRI.

$\text{MRI} = \frac{\text{number of (D+C+UP)}}{\text{N}}$, where D is Distinction, C is Credit , UP is Upper Pass awards and N is the population.

Higher class mean can result in higher MRI. The higher the MRI (closer to 1), the better the class performance. The MRI is expressed as a decimal. A school obtains a high MRI if the school mean is higher than the national mean, and the school standard deviation is small.

A high mean does not necessarily mean the school performed well, nor a small standard deviation means the school did well. It has got to be both, that is high mean and small standard deviation.This is linked to norm assessment than criterion assessment. Where a student mark is estimated as, $\text{Std.Mark(X)} = \frac{\text{P(A-B)}}{\text{C}} + \text{Q}$, where P-external sd, A-student internal mark, B –internal mean, C- internal sd, and Q is external mean.

Our discussion does not include student estimated mark, however, elements of variance is present, and the above is specific to school assessment. Businesses and industries interest vary and they make interpretations based on the type of data they have.

Calculating Mean and Standard Deviation of a Distribution

Mean or arithmetic mean is the sum of all scores divided by the number of scores. Deviation is the difference between the mean and a score. Standard deviation is the norm spread of scores from the mean score. The squared deviation is the variance.

Mean $\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$ or $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Example Calculate the mean of the data set: 21, 26, 32, 33, 33, 34, 35, 36, 39, 43

Solution

$$\bar{x} = \frac{21 + 26 + 32 + 33 + 33 + 34 + 35 + 36 + 39 + 43}{10}$$
$$= \frac{332}{10}$$
$$= 33.2$$
$$\approx 33$$

Standard Deviation $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

Take deviation of each score, square them then find the sum of the squares before dividing. So standard deviation is the square root of the mean deviation.

$$\text{mean deviation} = \frac{\sum (x_{ii} - \bar{x})^2}{n-1}$$

This should not be mixed up with the variance (s^2), which is the sum of all the squared deviations of all the data in the data set.

$$\text{variance} = s^2$$
$$s^2 = \sum (x_i - \bar{x})^2$$

This variance is often referred to as the second moment in statistics instead of the standard deviation. And we are using n – 1 instead of n in calculation of mean deviation, variance and standard deviation in order to avoid bias, as we are dealing with population sample. We divide by n when we deal with exact population.

Example Calculate the standard deviation of the data set: 21, 26, 32, 33, 33, 34, 35, 36, 39, 43

Solution

Variance

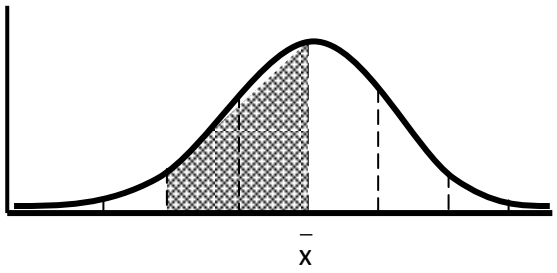
$$s^2 = \frac{\sum(x_i - \bar{x})}{n-1}$$
$$= \frac{\left[(21-33)^2 + (26-33)^2 + (32-33)^2 + 2(33-33)^2 + (34-33)^2 + (35-33)^2 + (36-33)^2 + (39-33)^2\right]}{10-1}$$
$$= \frac{\left[(-12)^2 + (-7)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2 + (3)^2 + (6)^2\right]}{9}$$
$$= \frac{[144+49+1+0+4+9+36]}{9}$$
$$= \frac{243}{9}$$
$$= 27$$

Standard Deviation (s)

$$s^2 = 27$$
$$s = \sqrt{27}$$
$$= 5.19612..$$
$$\approx 5$$

$\bar{x}$ -3s = 17, $\bar{x}$ -2s = 22, $\bar{x}$ -s = 27, $\bar{x}$ = 32, $\bar{x}$ + s =37 $\bar{x}$ +2s = 42, $\bar{x}$ +3s =47

68% of scores lie between $\bar{x}$ -s and $\bar{x}$ + s (27-37)
95% of scores lie between $\bar{x}$ -2s and $\bar{x}$ +2s (22- 42)
99.7% of scores lie between $\bar{x}$ -3s and $\bar{x}$ +3s (17- 47)



47.5% (34%+13.5%) of the scores lie between $\bar{x}$ - 2s and $\bar{x}$ (shaded region)
2.35% of the scores lie between $\bar{x}$ - 3s and $\bar{x}$ - 2s.

Since it is symmetrical about the mean, the same can be said about $\bar{x}$ + 2s and $\bar{x}$, and $\bar{x}$ + 3s and $\bar{x}$ + 2s.

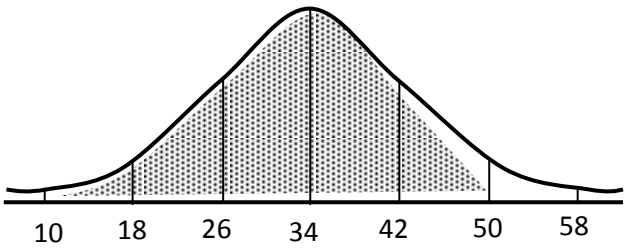
The process for calculating the standard deviation is long, so extending a frequency distribution table to include necessary columns will make calculation easy. As the data increases to 20 or more, calculation of variance and standard deviation becomes difficult especially without a scientific calculator, so use of table is very useful.

Frequency Distribution Table

Let us use Frequency Distribution table to find the mean and standard deviation of the school science examination marks.

Score x	Frequency f	Frequency x score fx	Deviation X - $\bar{x}$	Squared deviation $(x - \bar{x})^2$	Frequency x Squared deviation $f(x - \bar{x})^2$
11	1	**11**	**-23**	529	**529**
13	1	**13**	**-21**	441	**441**
16	1	**16**	**-18**	324	**324**
18	1	**18**	**-16**	256	**256**
19	2	**38**	**-15**	225	**450**
21	3	**63**	**-13**	169	**507**
22	1	**22**	**-12**	144	**144**
23	2	**46**	**-11**	121	**242**
24	2	**48**	**-10**	100	**200**
25	1	**25**	**-9**	81	**81**
26	3	**78**	**-8**	64	**192**
27	1	**27**	**-7**	49	**49**
28	2	**56**	**-6**	36	**72**
29	4	**76**	**-5**	25	**100**
31	2	**62**	**-3**	9	**18**
32	7	**224**	**-2**	4	**28**
33	7	**231**	**-1**	1	**7**
34	9	**306**	**0**	0	**0**
35	3	**105**	**1**	1	**3**
36	5	**180**	**2**	4	**20**
37	5	**185**	**3**	9	**45**
38	5	**190**	**4**	16	**80**
39	3	**117**	**5**	25	**75**
40	3	**120**	**6**	36	**108**
41	1	**41**	**7**	49	**49**
42	2	**84**	**8**	64	**128**
43	3	**129**	**9**	81	**243**
45	2	**90**	**11**	121	**242**
46	2	**92**	**12**	144	**288**
47	2	**94**	**13**	169	**338**
48	3	**144**	**14**	196	**588**
50	1	**50**	**16**	256	**256**
	∑f = 90	**∑fx =3031**		**$\sum f(x - \bar{x})^2$ =**	**6103**

μ = ∑fx ÷ ∑f = 3031/90 = 34 (whole) $\sigma = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{N}} = \sqrt{\frac{6103}{90}} = \sqrt{67.81} \cong 8$



Findings of school science result:

- The school mode is 34, median is 34 and the school mean is 34. The distributions **curve** is likely to be a normal curve since mean, mode and median are equal. The curve is symmetrical about the school mean.
- Skewness = 3(mean - median)÷ sd = 0; neither positive nor negative.
- Mean of Class A and Class B are below and above respectively about the school mean.
- Class B results influenced the school mean due to the high number of scores being above the mean of class A.
- There are several outliers that influenced standard deviation in the school analysis.
- $\sigma = 8$ is large for a maximum score of 50 with a mean of 34 for $+3\sigma$ to be equal to or below a score of 50 (highest possible student score).
- μ -3σ = 10 and μ$+3\sigma$ = 58 span beyond the set of data due to large σ of 8.
- Scores fall within 2σ above and 3σ below the mean of 34.

From these above findings, the head of the school assessment section can make recommendations to the particular subject teachers. This is to improve on teaching presentations and methods, inorder to increase student understanding of the concepts prescribed in the syllabus so the school can perform well in the following year.

A further investigation into the data with the use of z- score can be helpful. A z – score lies between -3 and +3 and is computed as $\mathbf{z} = \frac{\mathbf{x} - \bar{\mathbf{x}}}{\boldsymbol{\sigma}}$. A z-score of 0 means a student as scored a mean mark. A negative z – score (z < 0) means a student scored below the mean mark. And a positive z – score showed that the student scored above the mean mark.

Industries and businesses will have their own interpretations on such analyses however, the processes for computing the mean and standard deviation, and skewness is the same.

Skewness

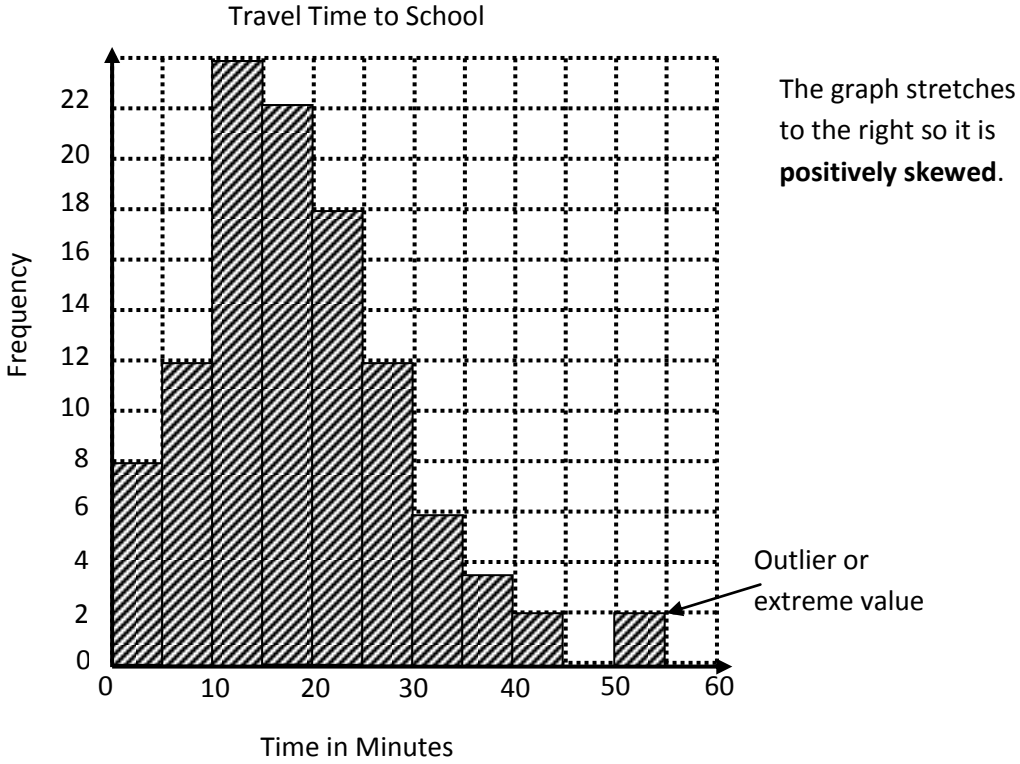
How can a distribution be normal distribution? It is just normal for a 1 year old to grasp, crawl, stand and eventually walk. Is it normal to see most of us especially adults crawling?

Majority of us have an average IQ and that is just normal. When we talk about normal distribution in statistics, we often have to deal with measures of skewness and kurtosis.

Skewness is a measure of symmetry, or more precisely, the lack of symmetry of a distribution about its mean.

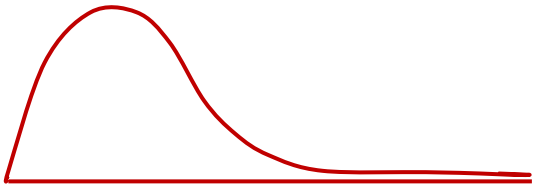
A distribution or data set is symmetric if it looks the same to the left and right of the centre point.

Below is a graph of St. Therese Primary School students travelling time to school recorded in 1914.



There are two types of skewness, the positively skewed and negatively skewed distribution. A **positively skewed** distribution has a "tail" which is pulled in the positive direction.

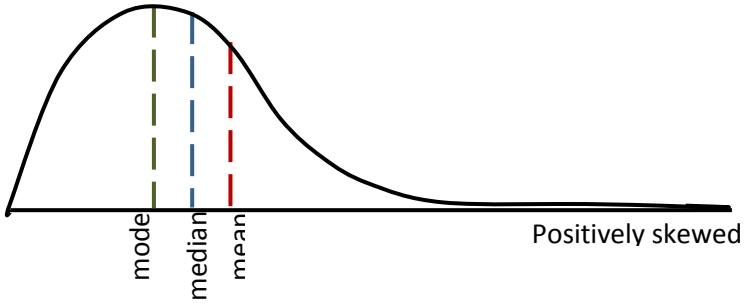
Example A spelling quiz intended for grade school pupils was given to senior high school students. It is expected that majority of the students will get high score or even a perfect score. The graphical presentation may appear like positively skewed as shown below.



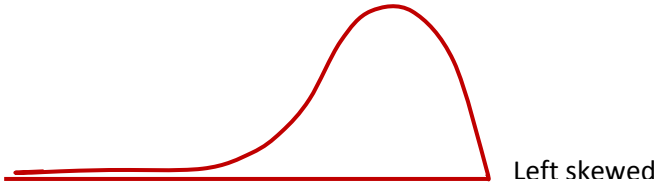
Right skewed

There seemed to be more scores on the left hand side, however the area below the curve on either side of the mean is the same. The mode is less than the mean.

The curve is positively skewed when the order is mode, median and mean. The long tail goes the positive direction.



On the other hand, if a spelling quiz intended for high school students was given to first graders, it is expected that majority of the pupils will fail and the graphical presentation may appear to be negatively skewed distribution.

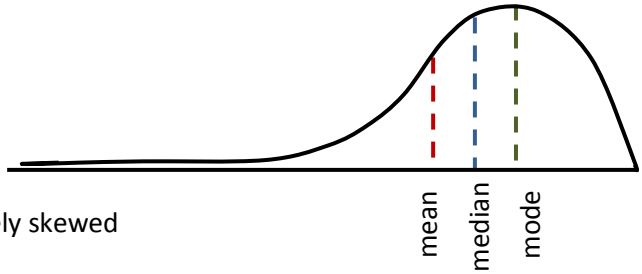


Left skewed

Notice that the graph does not look normal. There seemed to be more scores on the right hand side, however the area below the curve on either side of the mean is the same.

How can we produce a normal curve?

If we wish to have a normal curve, students must be assessed on what they are intended to answer. It is just normal that spelling quiz for high school is given to high school students and the like must be applied for grade schoolers.



Negatively skewed

The **Pearson's coefficient of skewness** is a measure of skewness denoted by ***SK*** and computed as:

$$SK_2 = \frac{3(\bar{x} - \hat{x})}{s} \quad \text{or} \quad SK_2 = \frac{3(\mu - \hat{\mu})}{s}$$

Note: The values of the Pearson's coefficient of skewness fall between -3 and +3. A value of SK which is closer to zero indicates that the distribution of the data set is symmetric.

Where μ = the population mean
$\bar{x}$ = the sample mean
$\tilde{x}$ = the sample median
s = the standard deviation

Example The following sets of data are the grade point average of a Statistics class during the first semester of the Academic Year 2015-2016. Compute the Pearson's Coefficient of Skewness.

2.9	3	3	3.1
2.4	2.5	2.8	2.8
2.1	2.1	2.2	2.3
1.8	1.9	1.9	2
1.5	1.5	1.1	1.2

Solution

To compute for the Pearson's second coefficient of skewness, compute first for the sample mean, sample median and sample standard deviation of the given data.

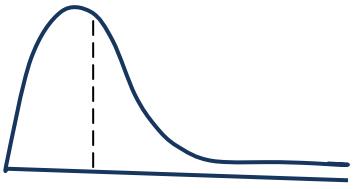
$\sum x = 44.1 \quad n = 20$

$\bar{x} = 2.21 \qquad \tilde{x} = 2.15 \qquad s = 0.61$

$$SK = \frac{3(2.21 - 2.15)}{0.61} \qquad SK = 0.2951$$

Conclusion

Since the computed SK value is 0.2951, this indicates that the data set shows a positively skewed distribution. The mean is on the right of mode.



Positively skewed

Note: If the computed SK is negative, it will show a negatively skewed distribution.

However, area before and after the cut line (mean) is the same.

Skewness can also be calculated by Pearson's first coefficient of skewness using mean and mode as

$$Sk_1 = \frac{\left(\bar{x} - \hat{x}\right)}{\sigma}$$

The above data is multimodal and too few so the Pearson's first coefficient of skewness is not applicable. We use Pearson's first coefficient of skewness if the data is unimodal and is large.

Without the curve, histogram, dot-plot or stem-and-leaf plot we cannot get a hint whether the distribution is normal. So we must note the following characteristics of the curve when we calculate the measures of central tendencies.

1. If $\bar{x} > \tilde{x}$ Right or positive skewed distribution,
2. If $\bar{x} < \tilde{x}$ Left or negative skewed distribution,
3. If $\bar{x} > \hat{x}$ Right or positive skewed distribution,
4. If $\bar{x} < \hat{x}$ Left or negative skewed distribution.

Example 2 A data set of 50 scores has a sum of 1258, and mode of 11 and a median of 23. Find the mean and state whether the distribution is negatively or positively skewed, and why.

Solution

Given $\sum x = 1258$, $n = 50$, $\hat{x} = 11$ and $\tilde{x} = 23$

$$\bar{x} = \frac{\sum x}{n}$$

$$= \frac{1258}{50}$$

$$= 25.18$$

25.18 > 11, therefore distribution is positively skewed.
25.18 > 23, therefore distribution is positively skewed.

You can use only one of the two characteristic as a reason for your statement. You do not have to use both.

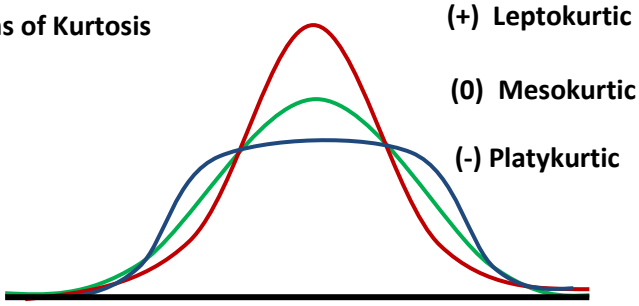
Kurtosis

The next figure shows a normal distribution that has something to do with skewness with normal curve. **Kurtosis** is incorrectly described as the degree of peakedness of a distribution and is usually taken relative to normal distribution.

A **normal distribution** is a **mesokurtic** distribution . A **leptokurtic** distribution has higher peak than the normal curve and a heavier tail; and a **platykurtic** distribution which is implied as flat topped (not always) has a lower peak than a normal distribution and has fewer and less extreme outliers than normal distribution.

Kurtosis is the measure of the dispersion of X around the two values $\mu \pm \sigma$, and two tails.

General forms of Kurtosis



When :

- $B_2 = 3$ mesokurtic,
- $B_2 < 3$ platikurtic,
- $B_2 > 3$ leptokurtic

Sample

$$B_2 = \frac{m_4}{(m_2)^2} = \frac{m_4}{\sigma^4}$$ Where m_2 and m_4 are respectively, second and fourth moments of the distribution. B_2 is the kurt [X]

In our discussion we will only consider four moments-mean, standard deviation, skewness and kurtosis. Physics applies fifth moment and upward.

Skewness and kurtosis is seldom discussed at this level in secondary school. However, the time is right that we discuss this because most data is never a normal distribution set. There always exist skewness and kurtosis.

Below are general forms of moments in statistics.
First moment (s = 1)

$$(x_1 + x_2 + x_3 + ... + x_n)/n)$$ $$\mu_i^{\,n} = \sum_{k=0}^{n} x_k^{\,i} P_k$$

Second moment (s = 2) sample variance

$$\sum(x_i - \mu_x)^2$$

Third moment (s = 3) Pearson's coefficients of skewness.

,.

$SK_1 = \frac{\bar{x} - \hat{x}}{s}$ (not to be used if the data is too few or there is an outlier),

$$SK_2 = \frac{3(\bar{x} - \tilde{x})}{s}$$

Fourth moment (s = 4)

$$\left(\frac{x_1^{4\backslash} + x_2^4 + x_3^4 + \ldots + x_n^4}{n}\right)$$

The rules for moments vary slightly when applied in matrix and geometry. But the general concept exists in different fields of mathematics.

The moments are found by $m = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^n$ for n = 2, 3, 4. Then take nth root. The first yields the standard deviation, while the other two are third and fourth moments. Use Pearsons skewness coefficients 1 and 2 if you have a problem with computing the skewness using **B_1 formula**. The answer you get as coefficient will be close but not exact, though the inference will be the same.

Let us find the skewness and kurtosis of the data from frequency distribution table on page 83. From the data, mean (μ)= 34 ,mode ($\tilde{x}$) = 34, median ($\hat{x}$) = 34, sd (σ) = 8 and kurt = 120 .

Skewness

$Sk_1 = (\bar{x} - \hat{x})/\sigma$

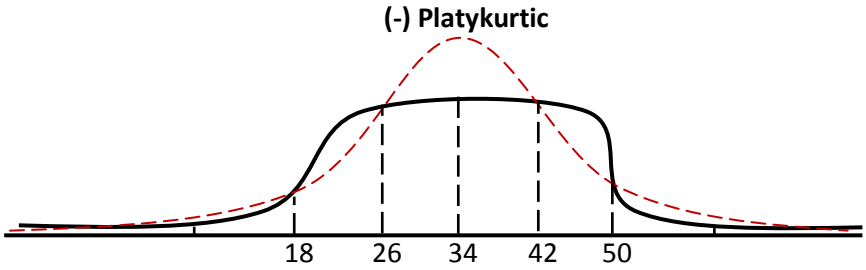
=(34 – 34)/8 = 0/8 = 0 Distribution is mesokurtic, most data (95%) lie within 2σ above and below the mean.

Kurtosis

$B_2 = m_4/(m_2)^2 = 120/8^4 = 120/4096 = 0.029296\ldots = 0.293$

$B_2 < 3$ or $B_2 - 3 = -2.707$

which is positive. The distribution is platykurtic, there is access data on the tails.



The skewness is 0 and the kurtosis is 0.293 with excess kurtosis -2.707 , and not 0. Therefore the data is platykurtic (B_2<3) with heavy tail on the left ($B_2 - 3 = -2.707$). That is realistic interpretation, as we do not expect any data above 50 (maximum score possible).

Since it is a school examination data, the items were easy to mosts students in the school (kurtosis of 0.293 means most data is around the mean), thus the mean was high. However, had weaker students need been attended to accordingly, the standard deviation would have been reduced (less than 8; excess kurtosis of -2.707) and mean could have increased (greater than 34).

A different statistical inference is made depending on the data source. We cannot say the same if the data generating such skewness and kurtosis is obtained from an industrial production line and businesses etc.

Take heed of the processes in calculating the four moments of statistics and then the kurtosis using the table.

x	**f**	**fx**	**$X - \bar{x}$**	**$f(x- \bar{x})^2$**	**$f(x- \bar{x})^3$**	**$f(x- \bar{x})^4$**
X_1						
X_2						
Xn						
Totals	∑f= n =	∑fx=		**$\Sigma f(x- \bar{x})^2=$**	**$\Sigma f(x- \bar{x})^3=$**	**$f(x- \bar{x})^4=$**
Means						
ROOTS						

Take note here that:

$n = \sum f$,

$\frac{\sum f(x - \bar{x})^2}{n - 1} = s^2$ is second moment

$\frac{\sum f(x - \bar{x})^3}{n - 1} = s^3$ is third moment,

We use s^2, s^3 and s^4 when we compute statistic of the sample. We use σ^2, σ^3 and σ^4 when we compute statistic of the population.

$\frac{\sum f(x - \bar{x})^4}{n - 1} = s^4$ is fourth moment.

Skewness

$B_1 = m_3/m_2^{3/2}$

$B_1 < 0$ negative skew, $B_1 = 0$ normal, $B_1 > 0$ positive skew

The formula can be used to find skewness only when the table as set in the previous page is completed to find moments. Otherwise, use Pearson's coefficients one or Pearson's coefficient two to find skewness.

Kurtosis

$B_2 = m_4/m_2^2$

$B_2 < 3$ or excess is less than zero is Platykurtic which has light tails , $B_2 = 3$ or excess is zero is Mesokurtic $B_2 > 3$ or excess is greater than zero is Leptokurtic with heavy tails.

Alternatively, since if the table is used the following formula can be used for population statistic:

$$a_4 = \sum \frac{(X_i - \overline{X})^4}{ns^4}$$

Recent published report by Dr. Westfall suggests that the mean (first moment) and variance (second moment) provide enough information and skewness and kurtosis are not really necessary.

One main reason to avoid the use of kurtosis is that there is a controversy over whether the kurtosis actually describes the peakedness and flatness; or the heaviness of the two tails relative to the rest of the distribution. There are also slight variations in both the formulae for skewness and kurtosis so it is important to state formula used in discussions of the distributions.

Learning Activity 11.3.2.5 and 11.3.2.6

 60 minutes

1. Which measure of central tendency is most appropriate to use in the following sets of data? Explain your answer in one sentence for each item.

   a. 23, 24, 27, 24, 28, 38, 32, 36, 35, 94

   b. 41, 44, 46, 44, 42, 37, 39, 40, 44, 37, 35, 44, 36, 44

   c. 87, 84, 79, 89, 86, 81, 83, 87, 91, 78, 85

   d. 4, 5, 6, 8, 12, 15, 18, 22, 25, 27, 28, 32, 98

   e. 56, 53, 65, 64, 67, 69, 73, 75

2. The following are the duration of calls (in minutes) made by a call centre agent in a four hour duty in a day.

   8, 4, 3, 5, 5, 6, 2, 3, 4, 8, 6, 5, 5, 3, 4, 2, 5, 6, 3, 4

Compute the Pearson's coefficient of skewness. Write whether it is positively or negatively skewed distribution.

3. Calculate skewness and kurtosis of the data set of a mathematics diagnostic test.

x	**f**	**fx**	$x-\bar{x}$	$f(x-\bar{x})^2$	$f(x-\bar{x})^3$	$f(x-\bar{x})^4$
3	1					
4	3					
5	5					
6	6					
7	8					
8	9					
9	5					
10	3					
Totals	**∑f=**	**∑fx=**		$\sum f(x-\bar{x})^2$ **=**	$\sum f(x-\bar{x})^3$ **=**	$\sum f(x-\bar{x})^4$ **=**
Means						
ROOTS						

11.3.2.7 Problems Involving Measures of Central Tendency

How is the measure of central tendency used in solving real-life problems and in making decisions? Your aim in this lesson is to apply what you have learned to real-life situations. The following problems will help you demonstrate your understanding on solving measures of central tendency.

Example 1 A student made a survey on the age(in years) of 20 mothers of grade 11 students. The following are the gathered data in random order.

34	37	53	47	39
45	40	36	31	46
51	42	41	38	36
35	39	48	45	41

Compute the (a) mean, (b) median and (c) mode and write your conclusion(interpretation) in 1 to 2 sentences.

Solution

a. MEAN $(\overline{X})$ =

$$\frac{34+37+53+47+39+45+40+36+31+46+51+42+41+38+36+35+39+48+45+41}{20}$$

MEAN $(\overline{X})$ = 41.1

The mean average age of parents (mothers) of grade 11 students is 41.1 or **41 years.**

b. MEDIAN $(\tilde{X})$

Arrange the data in ascending order.

31	34	35	36	36
37	39	39	38	40
41	41	42	45	45
46	47	48	51	53

To compute the median, we identify the 10^{th} and 11^{st} data since these are the middle data.

MEDIAN $(\tilde{X}) = \frac{40+41}{2}$ **MEDIAN $(\tilde{X})$ = 40.5**

The median average age of parents (mothers) of grade 11 students is 40.5 or **41 years.**

c. MODE ($\hat{X}$)

The data have four modes which are 36, 39, 41 and 45.

In this case, we may say that the type of average that will best describe and interpret the given data is the mean. We may say that the median is also appropriate but since there is no outlier, therefore we choose the mean. Definitely the mode cannot be considered as the best method because it appears to have four modes.

Example 2 The following is a table of weekly hours worked by 150 laborers in a certain project under the supervision of Mark Joseph.

Find the (a) Mean, (b) Median and (c) Mode.

Number of hours worked	Number of Laborers (*f*)
30-34	10
35-39	14
40-44	30
45-49	46
50-54	21
55-59	17
60-64	12

Solution

To compute the mean, we need to add two more columns for the class mark (X) and the product between the frequency and the class mark (*f*X).

Number of	Number of	X	*f* X
60-64	10	62	620
55-59	14	57	798
50-54	30	52	1560
45-49	46	47	2162
40-44	21	42	882
35-39	17	37	629
30-34	12	32	384
N = 150		∑*f*X=7035	

a. MEAN $(\overline{X}) = \frac{\Sigma fX}{N}$

MEAN $(\overline{X}) = \frac{7035}{150}$ **MEAN $(\overline{X})$ = 46.9**

The mean average weekly number of hours worked by 150 laborers in a certain project under the supervision of Mark Joseph is 46.9.

b. MEDIAN

To find the median, we add to more columns for Lower Class Boundary (LCB) and Cumulative Frequency Less Than (*Cf*<).

Number of hours worked	Number of Laborers (*f*)	X	*f* X	*Lb*	*Cf*<	
60-64	10	62	620	59.5	150	
55-59	14	57	798	54.5	140	
50-54	30	52	1560	49.5	126	
45-49	**46**	**47**	**2162**	**44.5**	**96**	**Median Class Modal Class**
40-44	21	42	882	39.5	50	
35-39	17	37	629	34.5	29	
30-34	12	32	384	29.5	12	

Locate the median class

$\frac{\Sigma f}{2}$ is the median class found at *cf* <

$\frac{\Sigma f}{2} = \frac{150}{2} = 75$ Locate 75 at CF<

The median class is located at 45 – 49 since 75 is located in CF< 96 for the cumulative frequency of that class ranges from 51 to 96.

From the median class, we identify the values of the following

$\frac{\Sigma f}{2} = 75$

$Lb_{\tilde{X}} = 44.5$

$f_{\tilde{X}} = 46$

$cf < = 50$ the cumulative frequency below (or less than) the median class

$i = 5$

Compute the median using the formula for grouped data

$$Median(\tilde{X}) = Lb_{\tilde{X}} + \frac{\frac{\Sigma f}{2} - Cf <}{f_{\tilde{X}}} (i)$$

$$\tilde{X} = 44.5 + \frac{75 - 50}{46}(5)$$

$$\tilde{X} = 44.5 + 0.54(5)$$

$$\tilde{X} = 44.5 + 2.72$$

$$\tilde{X} = 47.22$$

The median average weekly number of hours worked by 150 laborers in a certain project under the supervision of Mark Joseph is 47.22.

c. MODE

The median class happens to be the modal class because it has the highest frequency of 46. The modal class is the class with the highest frequency. Thus, the modal class is 45 – 49.

$Lb_{mo} = 44.5$ $i = 5$

$D_1 = fmo - f(class\ higher\ class)$ $D_1 = 46 - 30$ $D_1 = 16$

$D_2 = fmo - f(class\ lower\ class)$ $D_2 = 46 - 21$ $D_2 = 25$

$$Mode(\hat{X}) = Lb_{mo} + \frac{D_1}{D_1 + D_2}(i)$$

$$Mode(\hat{X}) = 44.5 + \frac{16}{16 + 25}(5) \qquad Mode(\hat{X}) = 44.5 + 1.95$$

$$Mode(\hat{X}) = 46.45$$

Thus, the mode of 46.45 falls within the class boundaries of 45 – 49 which is 44.5 – 49.5.

Learning Activity 11.3.2.7

20 minutes

The frequency distribution below shows the scores of 50 Grade 11 students in their Mathematics quarterly exam.

Scores (Classes)	Frequency
46 – 50	2
41 – 45	8
36 – 40	9
31 – 35	11
26 – 30	9
21 – 25	7
16 – 20	4

1. Complete the table by adding the necessary columns needed for computing the averages.

2. Compute the Mean, Median and Mode.

3. Compare the computed averages in one or two sentences.

Summative Task 11.3.2

60 minutes

A.Multiple Choice.Write the letter of your choice before each number.

_____1. What can be said on Vitz who obtained a score of 75 in a Grammar objective test?

a. He answered 75 items in the test correctly.
b. He answered 75% of the test items correctly.
c. His rating is 75.
d. He performed better than 25% of his classmates.

_____2. The summary of measures that divide a ranked data set into four equal parts.

a. Interquartile Range
b. Percentile
c. Quartile
d. Decile

_____3. The difference between the third quartile and the first quartile for a data set is Called

a. Quartile Deviation
b. Interquartile Range
c. Percentile Rank
d. Percentile

_____4. The following data give the price-earnings ratios of 12 companies.

18 16 38 20 20 18 34 7 58 31 19 22

What is the value of the 62^{nd} percentile?

a. 19 b. 20 c. 21 d. 22

_____ 5. Which of the measure of central tendency is the most reliable?

a. mean b. median c. mode d. fractiles

_____ 6. What other term is referred to as the extreme value in a distribution?

a. mean b. negative score c. outlier d. highest score

_____ 7. What measure of central tendency is most appropriate for the given data assuming that no further statistical treatment is needed?

2, 3, 3, 3, 3, 3, 3, 4, 5, 5, 5, 6, 7, 9, 9, 10, 10, 11, 11, 11, 13,14, 15, 16, 19, 22, 250

a. range b. mean c. median d. mode

_____ 8. Standard deviation is to variability as mode to _____.

a. level of difficulty c. correlation
b. discrimination d. central tendency

_____ 9. The grades of Jayvilyn in 5 quizzes are 85, 78, 82, 80 and 84. What must be her grade in the 6th quiz so that she will have an average of 83?

a. 83 b. 86 c. 87 d. 89

_____ 10. The mean of 7 quizzes of Janica is 85. If the lowest score of 80 is deleted, what will be her average?

a. 73.57 b. 85.00 c. 85.83 d. 99.17

B. Identification. Write the correct answer on the blank after each number.

1. It is the difference between the first quartile and the third quartile which is also called the midspread. _____

2. It refers to the degree of peakedness of a distribution _____

3. An effective graphical technique for showing both the skewness and kurtosis of data set usually presented through bar graphs

_____

4. A measure of symmetry, or more precisely, the lack of symmetry _____

5. A distribution whose "tail" is pulled in the right direction _____

6. A distribution that has a higher peak than the normal distribution and has heavier tails _____

7. A distribution that has a lower peak than a normal distribution and lighter tails

_____

8. What is also known as bell-shape or normal distribution? _____

9. A distribution, or data set, is said to be _____ if it looks the same to the left and right of the center point.

_____

10. A distribution that has a flat top near the mean rather than a sharp peak

_____

C. Below is the tabular presentation of the grades of 60 students in Statistics.

Classes	f
96 – 98	3
93 – 95	4
90 – 92	6
87 – 89	9
84 – 86	14
81 – 83	11
78 – 80	7
75 – 77	2
72 – 74	4

Compute:

a. Mean, Median and Mode

b. Range

c. Q_1

d. P_{60}

e. P_{32}

11.3.3: MEASURES OF SPREAD OR DISPERSION

Revisiting our example on page 45 we compared the mean, median, and mode of the scores in Science of the 3 students in their assignments.

Arvin	:	91	79	81	84	78	87	88
Amiel	:	46	59	61	84	85	86	167
Daniel	:	79	84	89	84	84	76	92

Solution

Name of students	MEAN	MEDIAN	MODE
Arvin	84	84	No mode
Amiel	84	84	No mode
Daniel	84	84	84

By observation, we can say that Arvin and Daniel's scores are more comparable than those of Amiel's scores. But the question is: "Who is more consistent between Arvin and Daniel in terms of their scores in Science?"

The lesson on measures of variability will tell us how the values (or scores) are scattered or clustered about the typical value (or the mean and median in this case).

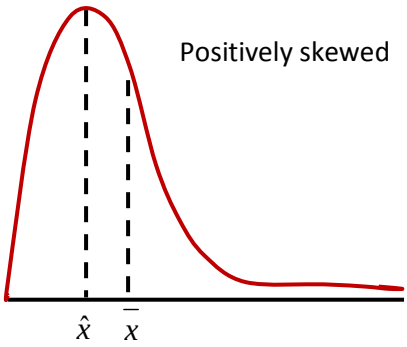
It is quite possible to have two sets of observations with the same mean or median but differs in the amount of spread or dispersion around the mean.

Smaller dispersion or variability of scores arising from the comparison often indicates more consistency and more reliability.

Measures of spread or dispersion refer to the variability (or spread) of the values about the mean.

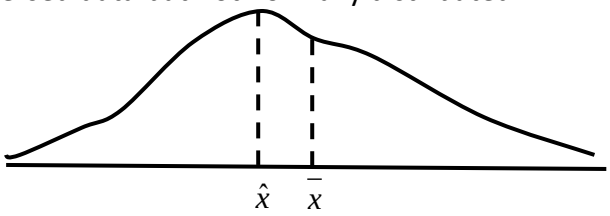
The measures of central tendency describe the most representative value of a group of data but it does not tell anything about the nature or the shape of the distribution whether the group is homogeneous (same or similar in nature) or heterogeneous (different in nature). The measures of spread or dispersion indicate the degree how variable the given data set are. The following are the different measures of spread.

Revisiting our lessons on skewness and kurtosis (pages 54-55), the following figures will give us an idea on how homogeneous and heterogeneous data are shown.

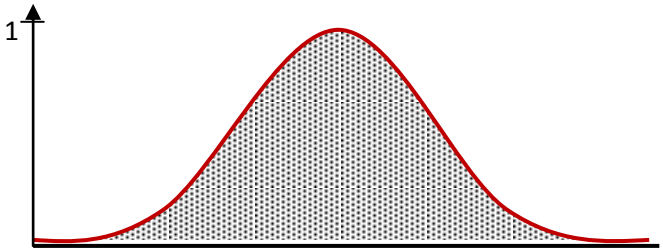


Since the graph shows a positively skewed distribution, this indicates that the data are homogeneous. Majority of the scores or data are **high**. Likewise for a negatively skewed distribution, majority of the score or data are **low**.

A normal distribution may indicate a dispersed data. The figure below will give us a clear view of a more dispersed data but not normally distributed.



A data set has a chance of being normally distributed only when the three averages (central tendencies) are the same or equal. And the skewness is 0, and kurtosis is 3. When the three conditions are met, the distribution can be said as a normal distribution. When it is a normal distribution, it can be called as a Gaussian Distribution with a probability density of 1.



The same probability curve can be used to illustrate z-score of students.

11.3.3.1 The Range

The **Range** is the simplest of measure of spread. It is the difference between the highest and lowest value of the distribution. It is a good measure of dispersion to use for small data sets but not advisable for large data set (or grouped data).

Range is the difference of the highest and the lowest data or score in the data set.

Range of Ungrouped Data

To find the range of the scores of the three students, we get the following differences.

Range = Highest Score – Lowest Score or **R = H. S – L. S**

Arvin's score	:	Range = 91 – 78	Range = 13
Amiel's score	:	Range = 167 – 46	Range = 121
Daniel's score	:	Range = 92 – 76	Range = 16

Based on the computed range as the simplest measure of dispersion, we may say that Amiel's scores are more spread (heterogeneous) while the computed range of Arvin's scores is the lowest which indicates that the scores of Arvin are clustered around the mean or more homogeneous than Daniel's score.

This means that Arvin is more consistent than Daniel. We will try to prove this in the succeeding lessons on variability.

Range of Grouped Data

In finding the range of a grouped data, it is calculated by getting the difference between the upper real limit (HRL) of the interval containing the largest score and the lower real limit (LRL) of the interval containing the lowest score. That is, we use the class boundaries to compute the range.

The URL is the upper class boundary of the highest class interval. And the LRL is the lower class boundary of the lowest class interval.

Range = Highest Real Limit – Lowest Real Limit
R = HRL – LRL

Say, if given is a class interval of 21-25, then the lower and upper class boundaries are 19.5 and 25.5 respectively. If the interval is the lowest class then we subtract 19.5 from the upper class boundary of the highest class intereval. If the interval is the highest, then we subtract the LRL from 25.5.

Example

Compute the range of the given frequency table on scores of students in their exam.

Class Interval	Frequency
91 - 95	1
86 - 90	2
81 – 85	3
76 – 80	6
71 – 75	7
66 – 70	4
61 – 65	5
56 – 60	3
51 – 55	2
46 – 50	2
N	35

Solution

Range = Highest Real Limit – Lowest Real Limit

R = HRL – LRL

= 95.5 – 45.5

Range = 50

11.3.3.2 Quartile Deviation or Semi-Interquartile Range

Revisiting our lesson about quartiles on page 43, we recall that quartiles divide the distribution by four (4). The three quartiles, Q_1, Q_2, and Q_3 are equal in terms of the proportion of observations on each portion.

We have proven that $Q_1 = P_{25}$, $Q_2 = P_{50}$, and $Q_3 = P_{75}$.

The interquartile range is the difference between the upper quartile and the lower quartile. The formula in finding the interquartile range for both ungrouped and ungrouped data is:

$IQR = Q_3 - Q_1$

The IQR is an improvement of the range because it eliminates the outliers (very high and very low extreme values).

The **Quartile Deviation or Semi-Interquartile Range** is a measure of spread focusing only in the middle 50% of the distribution or data scores.

The quartile deviation is more superior than the range and it may be used in sectioning or classifying pupils in a group. Since the focus is only in the middle 50% of the distribution, we can say that it is one-half the difference between the upper quartile and the lower quartile described by this formula.

$$Q.D = \frac{Q_3 - Q_1}{2}$$

Example

Compute the (a) Range and (b) Quartile Deviation of the distribution below.

Class Interval	Frequency	*Lb*	*Cf<*	
74 – 80	2	73.5	22	Q_3 class
67 – 73	**5**	**66.5**	**20**	
60 – 66	8	59.5	15	Q_1 class
53 – 59	**4**	**52.5**	**7**	
46 – 52	3	45.5	3	

Solution

a. Range = HRL – LRL
Range = 80.5 – 45.5
Range = 35

b. We compute first the Q_3 and Q_1 for grouped data.

$$Q_3 = Lb + \frac{-\frac{kN}{4} \; Cf <}{f_{Q_3}}(i)$$

Q_3 class = $\frac{kN}{4} = \frac{3(22)}{4} =$ **16.5**

Lb = 66.5
Cf< below = 15
f_{Q3} = 5
i = 7

$$Q_3 = 66.5 + \frac{16.5 - 15}{5}(7)$$

Q_3 = 66.5 + 2.1

Q_3 = 68.6

$$Q_1 = Lb + \frac{-\frac{kN}{4} \; Cf <}{f_{Q_1}}(i)$$

Q_3 class = $\frac{kN}{4} = \frac{1(22)}{4} =$ **5.5**

Lb = 52.5
Cf< below = *3*
f_{Q1} = 4
i = 7

$$Q_1 = 52.5 + \frac{5.5 - 3}{4}(7)$$

Q_1 = 52.5 + 4.375

Q_3 = 56.88

Quartile Deviation

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

$$Q.D. = \frac{68.6 - 56.88}{2}$$

QD = 5.86

The Quartile Deviation or Semi-Interquartile Range is 5.86.

LEARNING ACTIVITY 11.3.3.1 and 11.3.3.2

30 minutes

1. Calculate the range of the sets of ungrouped data below:

   (a) 2, 9, 3, 7, 8, 8, 3, 2, 4, 6, 5, 4, 3, 6

   (b) 25, 27, 29, 25, 29, 23, 26, 26, 26, 24

   (c) 19, 12, 17, 18, 13, 13, 13, 16, 18, 14, 15

2. Complete the table below by adding the necessary columns then compute the

   (a) Range and

   (b) Quartile Deviation or Semi-Interquartile Range.

Class Interval	Frequency
80– 84	3
75 – 79	4
70 – 74	6
65 – 69	9
60 - 64	10
55 – 59	11
50 – 54	5
45 – 49	2

3. The data below is the mathematical marks of a sample of 44 grade 11 students at a certain secondary school.

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

Calculate the Q_1, Q_2 and Q_3 and IQR

11.3.3.3 Average Deviation

The **Mean Average Deviation** of a set of data describe how much a set of data varies. Deviation is the distance between the measurement and the mean of the distribution.

Unlike the range and quartile deviation, the average deviation makes use of all the data in a distribution. Thus, it is a more reliable measure of spread.

The average deviation is the mean of all separate measures from the arithmetic mean but since the sum of the differences between each value and the arithmetic mean is always zero, the absolute values of the differences are summed up instead. The average deviation is an indicator of how compressed the group is on a certain measure. The average deviation can be computed as follows:

Average Deviation for UNGROUPED DATA

$$AD = \frac{\sum |X - \overline{X}|}{N}$$

Where X is the raw score

$\overline{X}$ is the mean

N is the number of data or scores

Average Deviation for GROUPED DATA

$$AD = \frac{\sum f|X - \overline{X}|}{N}$$

Where X is the midpoint or class mark

$\overline{X}$ is the mean

N is the number of data or scores

Example 1 (UNGROUPED DATA) Using the same example to test for consistency, let us compare the average deviation of Arvin and Daniel's scores in Science .

Arvin	:	91	79	81	84	78	87	88
Amiel	:	46	59	61	84	85	86	167
Daniel	:	79	84	89	84	84	76	92

Arvin's score (X)	Deviation $(X - \overline{X})$	Absolute Deviation $\lvert X - \overline{X} \rvert$
91	7	7
79	-5	5
81	-3	3
84	0	0
78	-6	6
87	3	3
88	4	4
$\Sigma X = 588$	**0**	$\Sigma \lvert X - \overline{X} \rvert = 28$

Daniel's score (X)	Deviation $(X - \overline{X})$	Absolute Deviation $\lvert X - \overline{X} \rvert$
79	-5	5
84	0	0
89	5	5
84	0	0
84	0	0
76	-8	8
92	8	8
$\Sigma X = 588$	**0**	$\Sigma \lvert X - \overline{X} \rvert = 26$

$\bar{X} = \frac{\Sigma X}{N}$ $\bar{X} = \frac{588}{7}$ $\bar{X} = \frac{\Sigma X}{N}$ $\bar{X} = \frac{588}{7}$

$\bar{X}$ **= 84** $\bar{X}$ **= 84**

$AD = \frac{\Sigma\left|X - \bar{X}\right|}{N}$ $AD = \frac{\Sigma\left|X - \bar{X}\right|}{N}$

$AD = \frac{28}{7}$ ***AD* = 4** $AD = \frac{26}{7}$ ***AD* = 6.5**

The Average Deviation of Arvin's scores is smaller than that of Daniel's scores. This is another indication that Arvin is more consistent than Daniel.

Example 2

The table below shows the gathered data by the National Statistics Office (NSO) about the literacy rate (literate only) of the people in a certain locality in 2014.

Age Group (Class Interval)	**Literate (Frequency)**	**X**	***fX***	**X -** $\bar{X}$	$\left\|X - \bar{X}\right\|$	$f\left\|X - \bar{X}\right\|$
61 – 70	78	65.5	5109	26	26	2028
51 – 60	96	55.5	5328	16	16	1536
41 – 50	125	45.5	5687.5	6	6	750
31 – 40	154	35.5	5467	-4	4	616
21 – 30	122	25.5	3111	-14	14	1708
11 – 20	83	15.5	1286.5	-24	24	1992
	N = 658		$\sum fX$ = 25989	0	90	$\Sigma f\left\|X - \bar{X}\right\|$ = 8630

$\bar{X} = \frac{\Sigma fX}{N}$ $\bar{X} = \frac{25989}{658}$ $\bar{X}$ = 39.50

$AD = \frac{\Sigma f\left|X - \bar{X}\right|}{N}$ ***AD* =** $\frac{8630}{658}$ ***AD* = 13.12**

The average deviation is 13.12.

11.3.3.4 The Standard Deviation of Ungrouped Data

The standard deviation is obtained by getting the square root of the mean of the squared deviations (from the mean) of a distribution.

> The **Standard Deviation** is a special form of measure of dispersion because it involves all the data scores in a distribution. It is the most important measure of homogeneity and heterogeneity of the distribution. **The standard deviation is the only measure of variability of distribution in making inferences.**

Distribution is the ascending or descending order set of a data showing the spread in a given limit.

Variability is the changes in the data set. Variance is the change in something. We use the term because X represents different values in a data set.

Dispersion is how spread each element of the data set is from one another. And deviation is the distance a data is positioned away from the mean of the data set.

The **variance** s^2 of Ungrouped Data is given by

$$s^2 = \left(\frac{x - \bar{x}}{N - 1}\right)^2$$

Standard Deviation for Ungrouped Data

$$s = \sqrt{\frac{\sum\left(x - \bar{x}\right)^2}{N - 1}}$$

Where S is the sample standard deviation
X is the raw score
$\bar{X}$ is the mean average
N is the number of data/cases

Note: We use N – 1 to make the sample variance an unbiased estimate of the population variance.

Example

Test for the consistency of scores between Arvin and Daniel.

Arvin	:	91	79	81	84	78	87	88
Daniel	:	79	84	89	84	84	76	92

Solution

Arvin's score (X)	Deviation (X - $\overline{X}$)	Squared Deviation $(X-\overline{X})^2$
91	7	49
79	-5	25
81	-3	9
84	0	0
78	-6	36
87	3	9
88	4	16
ΣX = **588**	**0**	$\sum\lvert X-\overline{X}\rvert$ **=144**

Daniel's score (X)	Deviation (X - $\overline{X}$)	Squared Deviation $(X-\overline{X})^2$
79	5	25
84	0	0
89	5	25
84	0	0
84	0	0
76	8	64
92	8	64
ΣX = **588**	**0**	$\sum\lvert X-\overline{X}\rvert$ **=178**

$$S=\sqrt{\frac{\Sigma(X-\overline{X})^2}{N-1}}$$

$$S=\sqrt{\frac{144}{7-1}}$$

$$S=\sqrt{24}$$

$$\boldsymbol{S = 4.90}$$

$$S=\sqrt{\frac{\Sigma(X-\overline{X})^2}{N-1}}$$

$$S=\sqrt{\frac{178}{7-1}}$$

$$S=\sqrt{29.6667}$$

$$\boldsymbol{S = 5.45}$$

Since the computed standard deviation of Arvin's scores is smaller than that of Daniel's, therefore Arvin's scores are more clustered around the mean. Thus, Arvin is more consistent than Daniel.

Note: If the standard deviation is smaller, the data are more homogenous and clustered about the mean. Otherwise, the data are heterogeneous and more dispersed.

LEARNING ACTIVITY 11.3.3.3 and 11.3.3.4

30 minutes

1. Given below are the scores of ten (10) Grade 11 students in their two quizzes in Algebra. Using the ungrouped data, calculate the following:

Quiz 1	22	18	16	24	9	10	19	23	15	27
Quiz 2	12	13	17	25	23	21	20	18	15	16

a. Range

b. Average Deviation

c. Standard Deviation

c. Based from the computed values, interpret the result in one to two sentences.

2. Calculate variance and standard deviation of the given ungrouped data by first completing the table.

Score	Frequency	Frequency x Score	Deviation	Squared Deviation	
x	**f**	**fx**	$x-\bar{x}$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
7	3	21			
8	5				
9	6				
10	10				
11	4				
12	3				
13	2				
14	1				
15	1				
	$\Sigma f=$	$\Sigma fx=$		$\Sigma f(x-\bar{x})^2=$	

11.3.3.5 The Variance

The variance is computed as the square of the deviations from the mean. It is simply the square of the computed standard deviation.

> The **Variance** is a measure of dispersion that considers the position of each observation relative to the mean.

Variance for Ungrouped Data

$$S^2 = \frac{\sum(X - \bar{X})^2}{N - 1}$$

Where S^2 is the variance for ungrouped data
X is the raw score
$\bar{X}$ is the mean average
N is the number of data/cases

Computing the variance between Arvin and Daniel's scores, we have

Arvin's computed Standard Deviation
$S = 4.90$

Variance $(S)^2 = (4.90)^2$
S = 24

Daniel's computed Standard Deviation
$S = 5.45$

Variance $(S)^2 = (5.45)^2$
S = 29.67

The average area that Arvin's scores are clustered about the mean is smaller than that of Daniel's.

Variance and Standard deviation for Grouped Data

For grouped data, the sample variance and standard deviation can be calculated using the following formula.

$$s^2 = \frac{\sum fX^2}{N-1} - \frac{\left(\sum fX\right)^2}{N(N-1)} \quad \text{and} \quad s = \sqrt{\frac{\sum fX^2}{N-1} - \frac{\left(\sum fX\right)^2}{N(N-1)}}$$

Where S = the standard deviation
S^2 = the variance
X = class mark or midpoint of each class
N = sample size

Example

The table below shows the anxiety level scores of 50 students in Mathematics. Compute the variance and standard deviation for grouped data.

Anxiety Level (Class Interval)	***f*** **(Frequency)**
60 – 65	3
55 – 59	6
50 – 54	7
45 – 49	8
40 – 44	9
35 – 39	5
30 – 34	4

Solution

Complete the table by adding the necessary columns needed for the formula.

Anxiety Level (Class Interval)	*f* (Frequeny)	*X* (Class Mark)	*fX*	X^2	fX^2
60 – 65	3	63	189	3969	11907
55 – 59	6	58	348	3364	20184
50 – 54	7	53	371	2809	19663
45 – 49	8	48	384	2304	18432
40 – 44	9	43	387	1849	16641
35 – 39	5	38	190	1444	7220
30 – 34	4	33	132	1089	4356
	N = 42		$\sum fX$=2001		$\sum fX^2$ = 98403

Step 1: Add the class mark column by getting the median for each class interval. It may also be obtained by dividing the sum of the lower limit and upper limit by 2.

For 60 – 65 class,

$$\frac{60+65}{2} = 63$$

Step 2: fX column is obtained by multiplying the frequency of each class and its corresponding class mark.

For 60 – 65 class
Frequency = 3
Class mark = 63

$fX = 3 \bullet 63$

$fX = 189$

Step 3: The X^2 column is just the square of each class mark.
For 60 – 65 class, the class mark is 63
$X^2 = 63^2 = 3969$

Step 4: The last column to be included is the fX^2 column. This is obtained by getting the product between the *frequency (f)* column and the square of the class mark (X^2) column.

For 60 – 65 class
Frequency = 3
$X^2 = 3969$

$fX^2 = 3 \bullet 63$

$fX^2 = 11907$

Step 5: Complete the table by getting the values of N, $\sum fX$, and $\sum fX^2$.These can be obtained the adding vertically the N, $\sum fX$, and $\sum fX^2$ columns.

N = 42 **$\sum fX$=2001** **$\sum fX^2 = 98403$**

Step 6: Use the Standard Deviation and Variance formula.
We just substitute the corresponding values obtained in steps 1 to 4 to find the Standard Deviation and Variance.

$$S = \sqrt{\frac{\sum fX^2}{N-1} - \frac{(\sum fX)^2}{N(N-1)}}$$

$$S = \sqrt{\frac{98403}{42-1} - \frac{(2001)^2}{42(42-1)}}$$

$S = \sqrt{2400.0732 - 2325.2038}$

$S = \sqrt{74.8694}$

$S = 8.65$

The standard deviation is 8.65

$$S^2 = \frac{\sum fX^2}{N-1} - \frac{(\sum fX)^2}{N(N-1)}$$

$$S^2 = \frac{98403}{42-1} - \frac{(2001)^2}{42(42-1)}$$

$S^2 = 2400.0732 - 2325.2038$

$S^2 = 74.8694$

The variance is 74.87

LEARNING ACTIVITY 11.3.3.5

30 minutes

1. Compute the variance and standard deviation for the ungrouped data.

data	frequency	Frequency x data	deviation	Squared deviation	Frequency x squared deviation
x	**f**	**fx**	**$x-\bar{x}$**	**$(x-\bar{x})^2$**	**$f(x-\bar{x})^2$**
11	2				
12	3				
13	5				
14	6				
15	9				
16	7				
17	3				
18	4				
19	1				
Total		$\sum fX =$			$\sum f(x-\bar{x})^2=$

2. The table below shows the ages of the employees in ABC Company.

Age in years (Class Interval)	**No. of employees (Frequency)**
51 – 55	8
46 – 50	17
41 – 45	22
36 – 40	16
31 – 35	15
26 – 30	3
21 – 25	2

Compute the variance and standard deviation for grouped data.

SUMMATIVE TASK 11.3.3

90 minutes

A. Multiple Choice. Write the letter of the correct answer before each number.

_____ 1. Which of the following refers to the difference between the highest and lowest values in a distribution?

a. range
b. variable
c. standard deviation
d. midpoint

_____ 2. Which of the following is the simplest measure of variation?

a. range
b. variance
c. average deviation
d. standard deviation

_____ 3. Which of the following is the most important and most widely used measure of dispersion?

a. average deviation
b. quartile deviation
c. standard deviation
d. coefficient of variation

_____ 4. The standard deviation is the most stable measure of dispersion because it is

a. most appropriately used when extreme scores exist.

b. expressed in percentage form.

c. the most commonly used indicator of the degree of dispersion and the most dependable estimate of variability.

d. the simplest measure of variation.

For items 5 to 15 refer to the following information.

Big Fruits Greenhouse was commissioned to develop an extra large mango for display in the Agri-Week Parade. A random sample of mangoes from each hybrid yielded these sizes (length) in centimetres for mature peak mangoes.

Hybrid A	8	9	11	15	16	16	18	20
Hybrid B	11	12	13	14	14	15	16	17

_____ 5. What is the mean average of Hybrid A?

a. 14.125 b. 15.5 c. 15 d. 16

_____ 6. What is the mean average of Hybrid B?

a. 14 b. 14.5 c. 14.125 d. 15

_____ 7. What is the range in Hybrid A?

a. 12 b. 16 c. 18 d. 20

_____ 8. What is the range in Hybrid B?

a. 6 b. 7 c. 8 d. 9

_____ 9. What is the variance of Hybrid A?

a. 2.0 b. 3.2 c. 4.0 d. 18.7

_____ 10. What is the variance of Hybrid B?

a. 2.0 b. 3.2 c. 4.0 d. 5.2

_____ 11. What is the standard deviation of Hybrid A?

a. 2.0 b. 3.59 c. 4.0 d. 4.32

_____ 12. What is the standard deviation of Hybrid B?

a. 2.0 b. 3.59 c. 4.0 d. 4.32

_____ 13. Which of the two data is clustered closely around the mean?

a. Hybrid A b. Hybrid B c. Both A and B d. Neither A nor B

_____ 14. Which is the best choice for displaying the big mangoes?

a. Use Hybrid A because it has the possibility of breeding 20cm mangoes.
b. Use Hybrid B because the breed is more consistent in its size.
c. Use Hybrid A since its average length of mangoes is greater than those of Hybrid B.
d. Grow another hybrid.

_____15. Which group has heterogeneous size of mangoes?

a. Hybrid A
b. Hybrid B
c. either A or B
d. Neither A nor B

B. The table below shows the age of employees in a university.

Class	**f**	**X**
55 - 60	4	
50 – 54	10	
45 – 49	15	
40 – 44	17	
35 - 39	12	
30 – 34	8	
25 - 29	2	
	N = 68	

Complete the table and compute the following:

1. Range

2. Quartile Deviation

3. Average Deviation

4. Variance

5. Standard Deviation

SUMMARY

- **Statistics** is the process of collection, presentation, analysis and interpretation of data.
- **Parameter** is the quantity calculated from the population or characteristic of a population.
- **Data** is factual information in the form of figures obtained from experiments or surveys.
- **Population** consists of the totality of the observation with which we are concerned.
- **Sample** is a subset of a population.
- **Observation** is any of information, whether it is numerical (quantitative) or categorical (qualitative) recorded.
- Distribution is the spread of data arranged in ascending or descending order.
- **Qualitative Data** are information that are descriptive in nature. It refers to attributes just like the information that we write in a curriculum vitae or bio-data.
- **Quantitative Data** are numerical information obtained through information or counting. Example: age, IQ scores, height, weight, etc.
- **Frequency Distribution** is a tabular arrangement of the data by categories showing the frequency of occurrence of values and classmarks or midpoints. It has a class frequency containing the number of class intervals.
- **Stem-and-leaf plots** is a method of organizing data in which the **stem values** or the leading digit for each observation are listed in a column and the **leaf values** or the trailing digit for each observation are then listed beside the corresponding stem.
- A **measure of central tendency** is a single, central value that summarizes a set of numerical data called **average.** It is used to describe what is 'typical' in a set of data. These are the **Mean**, **Median** and **Mode**.
- **MEAN** is the most common type of arithmetic average. The mean of the set of data is the sum of all the measurements divided by the number of measurements contained in the set of data. The symbol used to represent the sample mean average is $\overline{X}$ and population mean is μ. Thus $\mu = \frac{1}{n}\sum x_i$ and sample mean $\bar{x} = \frac{1}{n}\sum x_i$

- The **MEDIAN** is the middlemost value in a set of data arranged in ascending or descending order. The median is another type of average and most appropriate to use when the middle value is desired. The symbol used to represent the median is $\tilde{X}$.

- The **MODE** is the value or raw score which occurs most frequently in the set of data. It is the easiest type of average to compute and it can actually be found by inspection. The symbol used to represent the mode is $\hat{X}$.

- The **Percentiles** are the ninety-nine score points which divide the distribution into one hundred equal parts. It characterizes the values according to the percentage below the.

- **Quantile** is the division of items in a frequency distribution into groups such as decile, quartile and percentile.

- **Quartile**- each of the four groups into which a statistical sample may be divided; one of the three values that divide the total number of items in a frequency distribution into four groups, each containing a quarter of sample population. Q_1, Q_2, and Q_3 are called the 1st, 2nd, and 3rd quartiles respectively.

- **Decile** is one of the nine values that divide the total number of items in a frequency distribution into ten groups, each containing an equal number of items.

- **Cumulative frequency** of a data c is the sum of the frequencies of the data a, b and c. Each cumulative frequency is the sum of the frequency of the data and all the frequencies of the preceding data.

- **Outlier** is a data that is outside other values in a set of data.

- **Normal Distribution** is a probability frequency distribution for a random variable that theoretically takes on a bell shape and is symmetrical about the mean.

- **Ogive** is a cumulative frequency graph

- **68-75-99.7 Rule** is a Gaussian probability that 68% of the data lie within one standard deviation above and below the mean, 75% of the data lie within two standard deviations above and below the mean and 99.7% of the data lie within three standard deviations above and below the mean.

- **Skewness** is a measure of symmetry, or more precisely, the lack of symmetry. A distribution or data set is symmetric if it looks the same to the left and right of the center point. SK = $3(\bar{x} - \tilde{x})/\sigma$

- $z = \frac{x - \bar{x}}{\sigma}$

- Generally calculated as $\left(\frac{x_1^3 + x_2^3 + x_3^3 + ... + x_n^3}{n}\right)$

- For sample skewness $B_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(x - \bar{x}\right)^2\right]^{\frac{3}{2}}}$

- **Kurtosis** is the degree of peakedness of a distribution and is usually taken relative to normal distribution. $B_2 = \frac{m_4}{(m_2)^2}\frac{kurtosis}{variance^2}$

  Kurtosis is also defined as $a_4 = \sum \frac{\left(X_i - \bar{X}\right)^4}{ns^4}$ where

  N is sample size, X is i^{th} X value, $\bar{X}$ is the mean and s is the sample standard deviation.

- A **normal distribution** is a **mesokurtic (B=3)** distribution. A **leptokurtic (B>3)** distribution has higher peak than the normal curve with heavier tails and a **platykurtic (B<3)** distribution with assumed flat top, has a lower peak than a normal distribution with light tails. The excess in normal distribution is B-3 = 0 (since B = 3).

- **Measures of spread or dispersion** refer to the variability (or spread) of the values about the mean.

- The **Quartile Deviation or Semi-Interquartile Range** is a measure of spread focusing only in the middle 50% of the distribution or data scores.

- The **Standard Deviation** is a special form of measure of dispersion because it involves all the data scores in a distribution. It is the most important measure of homogeneity and heterogeneity of the distribution. The standard deviation is the only measure of variability of distribution in making inferences. $\sigma = \sqrt{\frac{1}{n}\sum\left(x_i - \bar{x}\right)^2}$. Often used is the formula $s = \sqrt{\frac{1}{n\text{-}1}\sum\left(x_i - \bar{x}\right)^2}$ if it is a sample data to avoid bias.

- The **Variance** is a measure of dispersion that considers the position of each observation relative to the mean calculated as $\sigma^2 = \frac{1}{n}\sum\left(x_i - \bar{x}\right)^2$ for population and $s_x^{\,2} = \frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2$ for the sample.

- A data is **homogenous** if it is clustered around the mean, but is **heterogenous** if it is spread away from the mean.

- **Pearsons Correlation Coefficient r is the formula**

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Where
n = the number of coordinate pairs
$\bar{x}$ = the mean of x values
S_x = the standard deviation of x values
$\bar{y}$ = the mean of y values
S_y = the standard deviation of y values

- **Least Squares Regression Line**

$y = a + bx$

$b = \frac{rs_y}{s_x}$

$a = \bar{y} - b\bar{x}$

Where
b = the slope
a = the y – intercept
r = Pearsons correlation coefficient
$\bar{x}$ = mean of x values
$\bar{y}$ = mean of y values
s_x = standard deviation of x values
s_y = standard deviation of y values

- **Discrete data** are variables that are unrelated and have a finite number of values collected by counting..

- **Continuous data** a variables that are uninterrupted collected by measuring.

- **Raw data** is data that is not organised in any way, often in ascending or descending order. A raw data is referred to as a distribution if it had been arranged in order or sequence.

ANSWER'S TO LEARNING ACTIVITIES

STUDENT LEARNING ACTIVITY 11.3.1.1 (*p. 9)*

A)

1. Qualitative
2. Quantitative
3. Quantitative
4. Qualitative
5. Qualitative
6. Qualitative
7. Quantitative
8. Quantitative
9. Quantitative
10. Qualitative

B)

1. Continuous
2. Discrete
3. Continuous
4. Continuous
5. Discrete
6. Continuous
7. Continuous
8. Continuous

STUDENT LEARNING ACTIVITY 11.3.1.2 and 11.3.1.3 ***(p. 22 – 24)***

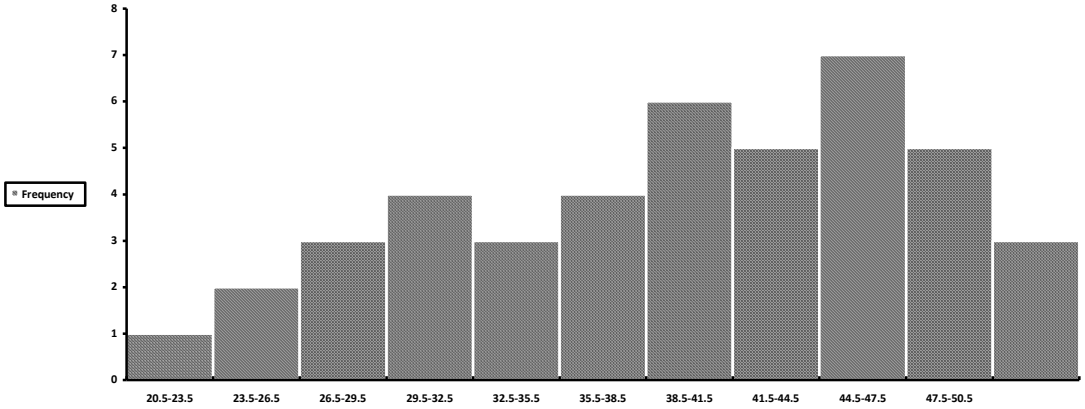
A. 1. D
2. B
3. C
4. A
5. B
6. D
7. C

B. 1.R = 27 $i = 3$

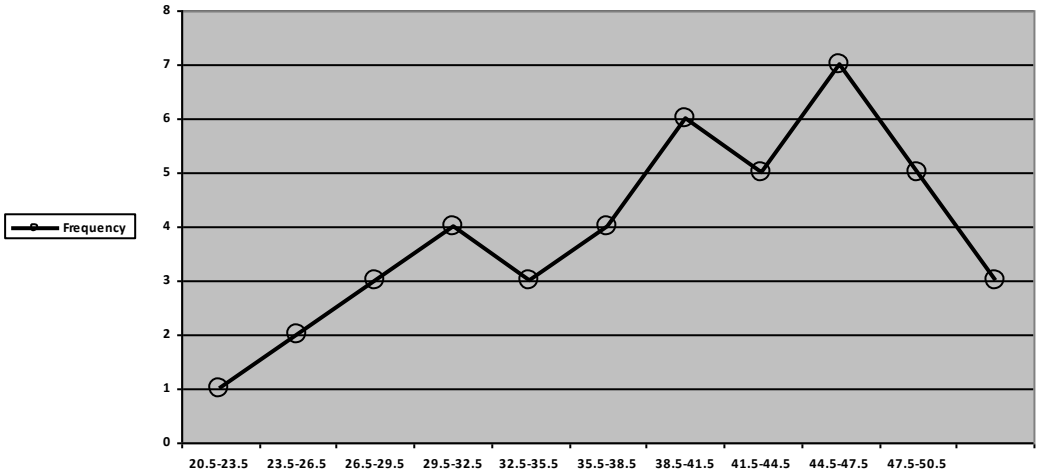
Classes	f	X
48 - 50	5	49
45 - 47	7	46
42 - 44	5	43
39 - 41	6	40
36 - 38	4	37
33 - 35	3	34
30 - 32	4	31
27 - 29	3	28
24 - 26	2	25
21 - 23	1	22

N = 40

2. Histogram



Frequency Polygon



3.

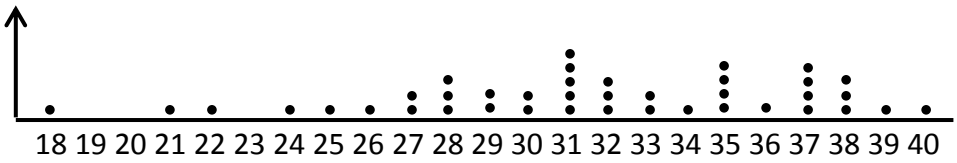
Class Group	Class centre (X)	f	Cf
11-15	13	1	1
16-20	18	0	1
21-25	23	2	3
26-30	28	4	7
31-35	33	6	13
36-40	38	7	20
41-45	43	10	30
46-50	48	10	40
		$\Sigma f=40$	

STUDENT LEARNING ACTIVITY 11.3.1.4 *(p. 27)*

1.

Stem	Leaf
1	8
2	1, 2,4, 5, 6, 7, 7, 8, 8, 8, 9, 9
3	0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 5, 6, 7, 7, 7, 7, 8, 8, 8, 9
4	0

2.Dotplot

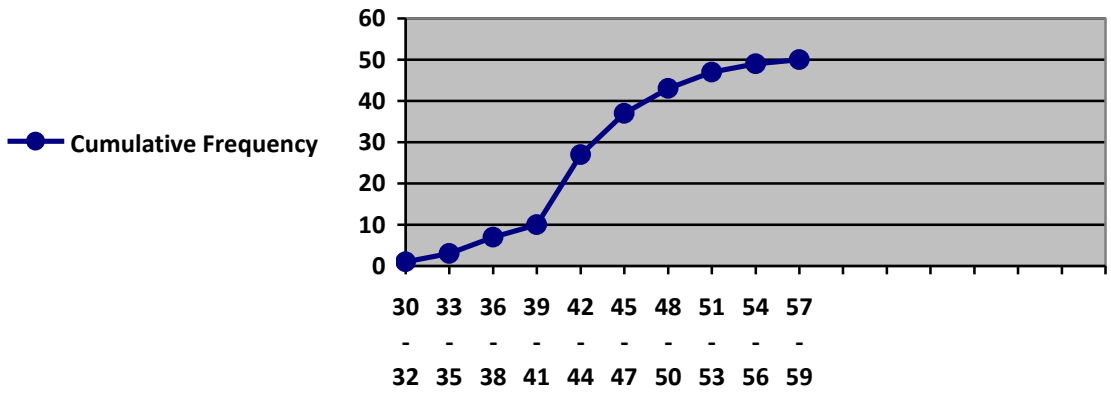


STUDENT LEARNING ACTIVITY 11.3.1.5, 11.3.1.6 and 11.3.1.7 (*p 37-39)*

1.

Classes	*f*	*Cf<*	*Rf*
57 – 59	1	50	2%
54 – 56	2	49	4%
51 -53	4	47	8%
48 – 50	6	43	12%
45 – 47	10	37	20%
42 – 44	17	27	34%
39 – 41	3	10	6%
36 – 38	4	7	8%
33 – 35	2	3	4%
30 – 32	1	1	2%
			100%

b. Construct a cumulative frequency less than or OGIVE.



2. n = 8, $\Sigma x = 80$, $\Sigma y = 190$, $\bar{x} = 10$, $\bar{y} = 23.75 \approx 24$

Slope = (24 - 15)/(10 - 4)= 9/6 = 1.5
When y = 1.5x + b at (10,24) we get 24 = 1.5(10) + b, therefore b = 9

Equation y = 9 + 1.5x

3. n = 8, $\Sigma EM = 405$, $\Sigma E = 58$ $\Sigma M = 58$, $\Sigma E^2 = 454$ $\Sigma M^2 = 444$ $(\Sigma M)^2 = 3364$

$$b = \frac{N\sum EM - \left(\sum E\right)\left(\sum M\right)}{N\sum E^2 - \left(\sum E\right)^2} = \frac{10*405 - 58*58}{10*454 - 3364} = 0.6 \text{ and}$$

$$a = \frac{\sum M - b\sum E}{N} = \frac{58 - 2.4*58}{10} = 2.4$$

M= 2.4 + 0.6E

Summative Assessment 11.3.1 (*P. 40-46)*

A.
1. a
2. b
3. d
4. a
5. c
6. b
7. a
8. b
9. d
10. a
11. b
12. c
13 c
14. b
15. a
16. d
17. b
18. c
19. b
20. b

B. 1. Frequency Distribution table

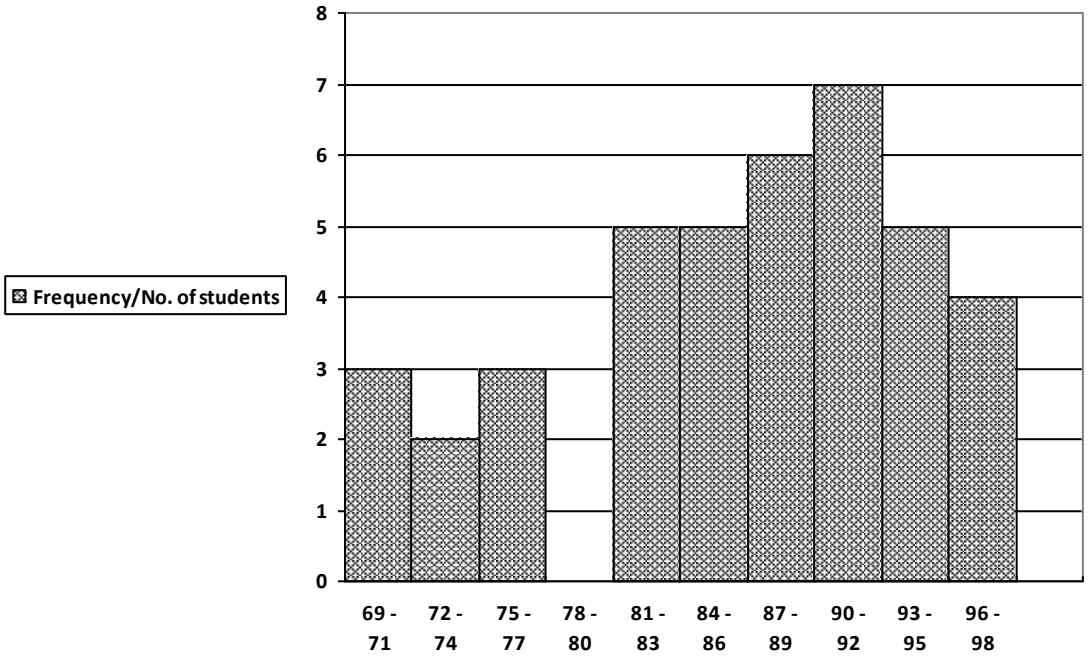
Test score (X)	Tally	f	fX
1	/	1	1
2	//	2	4
3	//	2	6
4	𝍸 𝍸 /	11	44
5	𝍸 𝍸	10	50
6	𝍸 𝍸 //	12	72
7	𝍸 //	7	49
8	𝍸	4	32
9	/	1	9
	Σf, ΣfX	50	267

2. Frequency Distribution Table

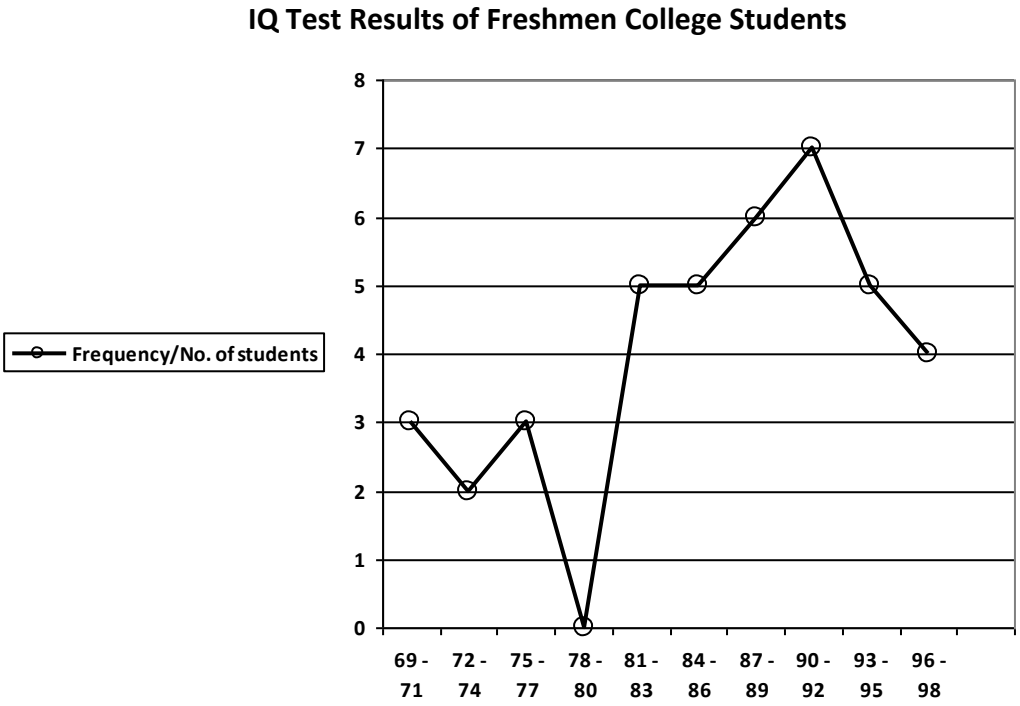
IQ Test Results of Freshmen CollegeStudents

Classes	f	X	Cf<	RF%
96 - 98	4	97	40	10
93 - 95	5	94	36	12.5
90 - 92	7	91	31	17.5
87 - 89	6	88	24	15
84 - 86	5	85	18	12.5
81 - 83	5	82	13	12.5
78 - 80	0	79	8	0
75 - 77	3	76	8	7.5
72 - 74	2	73	5	5
69 - 71	3	70	3	7.5
	N = 40			100

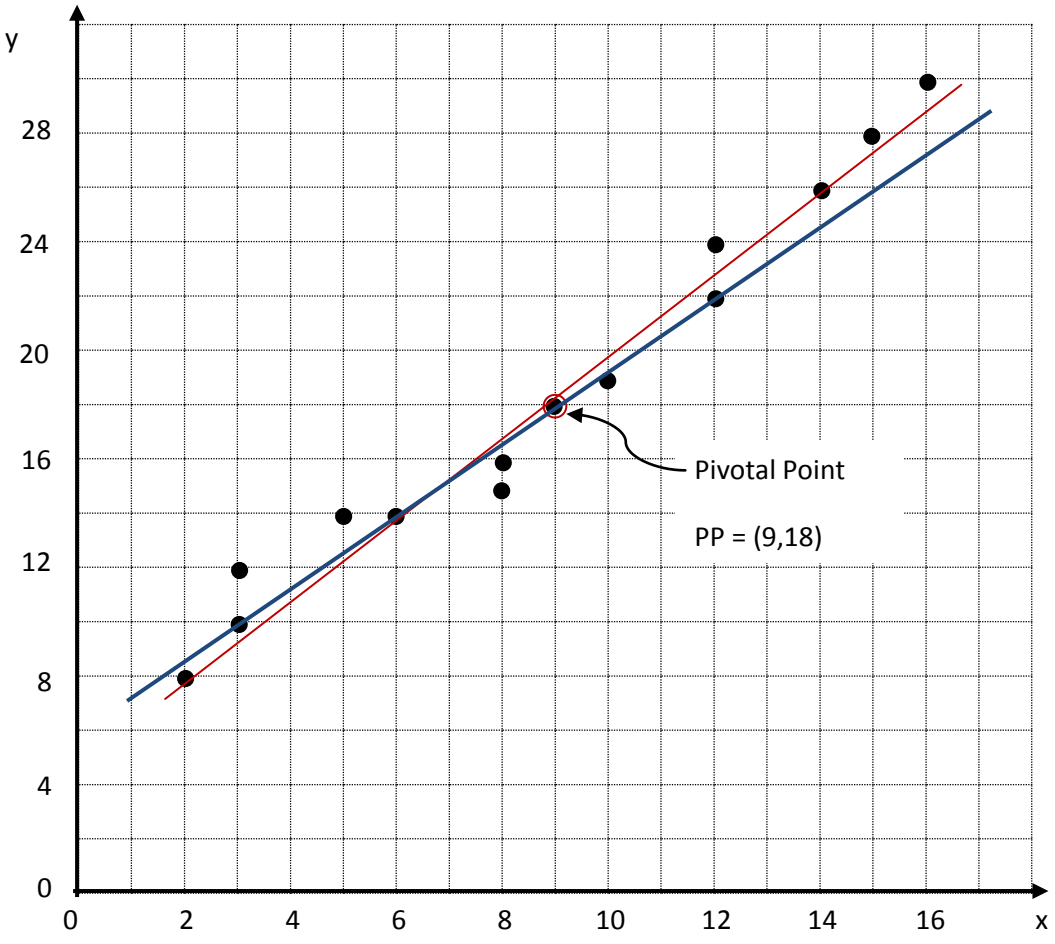
Histogram **IQ Test Results of Freshmen College Students**



Frequency Polygon



3.Scattergram



Using pivotal point y = 4/3x + 6

4. $\Sigma x = 42$ $\Sigma y = 107$ $\Sigma x^2 = 364$ $(\Sigma x)^2 = 1764$ $\Sigma xy = 518$ $(\Sigma x)(\Sigma y) = 4494$
Y = 41 -3.3X

STUDENT LEARNING ACTIVITY 11.3.2.1, 11.3.2.2 and 11.3.2.3 *(p. 57-58)*

1. a. Mean = 38.56
Median = 42
Mode = 43

b. Mean = 23.11
Median = 24
Mode = 22 and 25 Bimodal

c. Mean = 85.71
Median = 86
Mode = No mode

2. a.

Score	*f*	*X*	*fX*	*Cf<*
59 – 62	3	60.5	181.5	57
55 – 58	4	56.5	226	54
51 – 54	6	52.5	315	50
47 – 50	8	48.5	388	44
43 – 46	11	44.5	489.5	36
39 – 42	10	40.5	405	25
35 – 38	8	36.5	292	15
31 – 34	5	32.5	162.5	7
27 – 30	2	28.5	57	2
	Σf =		ΣfX = 2516.5	

b. Mean = 44.15
Median = 43.59
Mode = 43.5

c. Mean is greater than the mode and median.

d. Median as the class groups do not specify the data per class.

STUDENT LEARNING ACTIVITY 11.3.2.4 (*p. 70-71)*

1. a. $P_{80} = 438$
   b. $P_{14} = 360$
   c. $Q_1 = 380$
2. a. P_{60} class = 7001 – 8000 (24^{th} score)

P_{60} = 7000.5 + [((60x40/100) -12)/9] x = 7000.5 + 12/9 x 999
= 7000.5 +123 = 7123.5

c. Q_3 class =3(40+1)/4 = 8001 – 9000 (30.75^{th} score)

Q3 = 8000.5 + [(3 x 40/4 - 26)/8] x 999 = 8000.5 + 0.5 x 999 = 8000.5 + 499.5
= 8500

STUDENT LEARNING ACTIVITY 11.3.2.5 and 11.3.2.6 *(p 94 - 95)*

1.
   a. median
   b. mode
   c. mean
   d. median
   e. mean

2. Mean = 4.55
   Median = 5
   Standard Deviation = 1.7
   Skewness = - 0.79 It is negatively skewed distribution.

3.

x	f	fx	$x-\bar{x}$	$f(x-\bar{x})^2$	$f(x-\bar{x})^3$	$f(x-\bar{x})^4$
3	1	3	-3.975	15.800625	-62.807484375	249.6597504
4	3	12	-2.975	26.551875	-78.99182813	183.0832038
5	5	25	-1.975	19.503125	-38.51867188	76.07437695
6	6	36	-0.975	5.70375	-5.56115625	0.000542213
7	8	56	0.025	0.005	0.000125	0.000003125
8	9	72	1.025	9.455625	9.692015625	9.93431602
9	5	45	2.025	20.503125	41.51882813	84.07562695
10	3	30	3.025	27.451875	83.04192188	215.2018137
Totals	**∑f=40**	**∑fx=279**	$\Sigma(x-\bar{x})$ = -3.8	$\sum f(x-\bar{x})^2$ = 124..975	$\sum f(x-\bar{x})^3$ = -51.62625	$\sum f(x-\bar{x})^4$ = 818.0296332
Means		6.975	0.095	3.204487	-1.32375	20.9751188
ROOTS				1.790	-1.097999	2.140061

Skewness = -1.098 (tail to the left)
Kurtosis = 2.14 is < 3 so not Normal distribution
Excess = - 0.86 on the left

STUDENT LEARNING ACTIVITY 11.3.2.7 ***(p. 100)***

1.

Scores (Classes)	Frequency	*X* (Class Mark)	*fX*	*Cf<*
46 – 50	2	48	96	50
41 – 45	8	43	344	48
36 – 40	9	38	342	40
31 – 35	11	33	363	31
26 – 30	9	28	252	20
21 – 25	7	23	161	11
16 – 20	4	18	72	4
	N = 50		$\sum fX$ = 1630	

2. Mean = 32.6
   Median = 32.77
   Mode = 33.11

3. Answers may vary.

   Possible answer:
   Since the mean and median differ only by 0.17, these two averages can be considered more accurate than the mode. As regards the frequency column, the distributions seem to be normal. Thus, the mean average is the most appropriate type of average to use and most reliable measure of central tendency.

Summative Assessment 11.3.2 ***(p. 101 - 103)***

A.
1. a
2. c
3. a
4. b
5. a
6. c
7. d
8. d
9. d
10. c

B.
1. Quartile deviation
2. kurtosis
3. histogram
4. skewness
5. positive skew
6. leptokurtic
7. platykurtic
8. mesokurtic
9. symmetrical
10. platykurtic

C. Below is the tabular presentation of the grades of 60 students in Statistics.

Classes	f	X	fX	Cf<
96 – 98	3	97	291	60
93 – 95	4	94	376	57
90 – 92	6	91	546	53
87 – 89	9	88	792	47
84 – 86	14	85	1190	38
81 – 83	11	82	902	24
78 – 80	7	79	553	13
75 – 77	2	76	152	6
72 – 74	4	73	292	4
	N = 60		5044	

a. Mean = 84.1
Median = 85
Mode = 85

b. Range = 27

c. $Q_1 = 81.32$

d. $P_{60} = 85.21$

e. $P_{32} = 82.19$

STUDENT LEARNING ACTIVITY 11.3.3.1 and 11.3.3.2 (*p. 110 - 111)*

1. A. range = 7
B. Range = 6
C. Range = 7

2.

Class Interval	Frequency	Cf<	
80– 84	3	50	
75 – 79	4	47	
70 – 74	6	43	Q_3 class
65 – 69	9	37	
60 - 64	10	28	
55 – 59	11	18	Q_1 class
50 – 54	5	7	
45 – 49	2	2	

a. Range = 40

b. Quartile Deviation
Q3 = 69.92 Q1 = 57
Quartile Deviation or Semi-Interquartile Range = 6.46

3.

Score	Frequency	Cf<
15	2	44
14	4	42
13	5	38
12	10	33
11	5	23
10	18	18

$Q_1 = 10$
$Q_2 = 11$
$Q_3 = 12$
IQR = 2

STUDENT LEARNING ACTIVITY 11.3.3.3 to 11.3.3.4 (*p. 116 - 117)*

1.

		Quiz 1	**Quiz 2**
A.	range =	18	13
B.	Average deviation =	4.7	3.4
C.	Standard deviation =	5.93	4.24
D.	Interpretation:		

2.

Age in years (Class Interval)	No. of employees (Frequency)	X	fX	X- $\overline{X}$	F(X- $\overline{X}$)^2
51 – 55	8	53	424	12	1152
46 – 50	17	48	816	7	833
41 – 44	22	43	946	2	88
36 – 40	16	38	608	-3	144
31 – 35	15	33	495	-8	960
26 – 30	3	28	84	-13	507
21 – 25	2	23	46	-18	648
	N = 83		∑fX = 3419		∑fX^2 = 4332

$\overline{X}$ =ΣfX/N = 41

a. Variance = ∑fX/N = 52.19

b. Standard Deviation = $\sqrt{(\sum fX /N)^2}$ = 7.22

STUDENT LEARNING ACTIVITY 11.3.3.5 (*p. 121)*

1.

Data	Frequency	Frequency x data	Deviation	Squared deviation	Frequency x squared deviation
x	**f**	**fx**	**x-** $\overline{x}$	**(x-** $\overline{x}$ **)^2**	**f(x-** $\overline{x}$ **)^2**
11	2	22	-3.9	15.21	30.42
12	3	36	-2.9	8.41	25.23
13	5	65	-1.9	3.61	18.05
14	6	84	-0.9	0.81	4.86
15	9	135	0.1	0.01	0.09
16	7	112	1.1	1.21	8.47
17	3	51	2.1	4.41	13.23
18	4	72	3.1	9.61	38.44
19	1	19	4.1	16.81	16.81
Total	∑f = 40	$\sum$ fX = 596	∑(**x-** $\overline{x}$ **) =** 0.9	**∑(x-** $\overline{x}$ **)^2=** 60.09	**∑ f(x-** $\overline{x}$ **)^2=** 155.6

$\overline{x}$ = 14.9

Variance S^2 = ∑ **f(x-** $\overline{x}$ **)^2/**∑f = 155.6/40 =3.89

Standard deviation S = 1.97

2.

Age in years (Class Interval)	No. of employees (f)	Class Centre (X)	fX	$(X - \bar{X})$	$f(X - \bar{X})^2$
51 – 55	8	53	424	12	1152
46 – 50	17	48	816	7	833
41 – 45	22	43	946	2	88
36 – 40	16	38	608	-3	144
31 – 35	15	33	495	-8	960
26 – 30	3	28	84	-13	507
21 – 25	2	23	46	-18	648
Total	**83**		**3419**		**3499**

$\bar{X} = 41$ Variance (s^2) = 42.7 Standard Deviation (s) = 6.5

Summative Assessment 11.3.3 ***(P.122- 125)***

1. a
2. a
3. c
4. c
5. a
6. a
7. a
8. a
9. d
10. c
11. d
12. a
13. b
14. b
15. a

B. 1. Range = 36
2. Quartile Deviation = 11.08
3. Average Deviation = 6.10
4. Variance = 56.16
5. Standard Deviation = 7.49

REFERENCES

Aguinaldo, E. Et. al. (2011). ***Beginning Statistics***. TCS Publishing House, Plaridel Bulacan, Philippines

Bluma, Allan G. (1992). ***Elementary Statistics: A Step by Step Approach.*** *Iowa*. WM. C. Brown

Campena, Francis Joseph H. (2009). ***High School Statistics***. DIWA Learning Systems Inc. 1229 Makati City, Philippines

Coolidge, F. (2006). ***Statistics: A Gentle Introduction***, 2nd Edition. SAGE Publications, Inc. 2455 Teller Road Thousand Dacks, California 91320

Devore, et. al. (2005). ***Statistics: The Exploration and Analysis of Data***. Thompson Learning, Inc. USA

Glencoe/McGRAW-HILL. (1986). ***Mathematics Skills for Daily Living.*** USA. Laidlaw Brothers, Publishers.

Hogg (2005). ***Introduction to Mathematical Statistics***. Pearson Education, Inc. Upper Saddle River, New Jersey 07458

Navidi, W. (2010). ***Principles of Statistics for Engineering and Scientists***. McGraw-Hill Co., Inc. 1221 Avenue of the Americas, NY,NY 10020

Reyes, Milagros Z.(1998). ***Statistical Methods in Education***. REX Bookstore Inc., Manila, Philippines

Villamoran, E. Et. al. (2010). ***Statistics***. TCS Publishing House, Plaridel Bulacan, Philippines

http://www.slideshare.net/maggiev/the-interpretation-of-quartiles-and-percentiles

http://www.mathsisfun.com/data/quartiles.html

http://www.mathsisfun.com/data/percentiles.html

http://www.icoachmath.com/problems/problemslink.aspx