

CHAPTER 4: STATISTICS AND DATA MANAGEMENT

LEARNING OBJECTIVES

1. Define Statistics and Describe the range of applications of statistics.
2. Determine and assess the different types of Descriptive Statistics.
3. Define Data and Data Management
4. Construct and Organize data using graphs and charts and evaluate data using the different types of statistical tools.

4.1 STATISTICS

- Statistics is the science of collecting, describing, interpreting data and making decisions based on data.
 - For instance, collecting numerical facts and figures about the number of people tested positive for Corona virus disease 2019 (COVID-19), then describing the possibility of how they got infected, tracing other people who may be exposed to them and may also be infected and in the end would come up with list of recoveries, deaths and those that needs to be quarantined. Statistics showed that a pandemic of respiratory disease is spreading from person to person caused by novel corona virus.
 - Statistics is so indispensable that bankers used it to estimate the number of clients who will be making deposits as compared to the number of clients requesting for loans, businesses rely on census data to determine the population in a given area, ages, income, educational attainment, occupations, etc. of possible customers.
 - It is a professional tool to systematically organize data, to reveal and quantify hidden patterns from chaos, to summarize or to categorize variations from population, to investigate complicated human behaviors, to make prediction based on existing knowledge, to review and to assess existing technologies and methods and to conceptualize new ideas which may lead to improvements.
 - Statistics is a branch of applied mathematics its applications can also be found in integration, differentiation, and algebra.

4.2 BRANCHES OF STATISTICS

1. Descriptive Statistics

Descriptive statistics is the term given to the analysis of data that helps describe, show, or summarize data in a meaningful and useful way. For example, the GPA of students in a class can be listed from highest to lowest for every subject and the students can be categorized accordingly.

The process of descriptive statistics can be carried out using the measures of central tendency and measures of spread. In this process it is useful to summarize group of data using tables, graphs, and statistical analysis.

2. Inferential Statistic

Inferential statistics is used to make predictions or comparisons about a larger group (population) using information gathered about a small part of population. It is produced through complex mathematical calculations that scientist used to infer.

Inferential statistics is generally used when the user needs to make a conclusion about the whole population at hand, and this is done using the various types of tests such as Linear Regression Analysis, Analysis of Variance, Analysis of Co-variance, Statistical Significance (T-Test) and Correlation Analysis.

4.3 DATA MANAGEMENT

Data refers to information coming from observations, counts, measurements, or responses. They are collection of facts and figures and individual pieces of information recorded and used for the purpose of analysis. Data are raw information from which statistics are created.

Example 1: Information from the profile of students like age, course, gender, year level, income of parents, etc. Another example are consumers preferences such as their soap, shampoo, toothpaste, lotion, etc.

Data management is a process that includes acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data for its users.

Example when making business decisions, commercial and industrial companies rely on big data to gain deep insights into customer behavior, trends, and opportunities for creating extraordinary customer experiences.

4.4 DATA COLLECTION

Data collection is defined as the procedure of collecting, measuring, and analyzing accurate insights for research using standard validated techniques.

The objective of data collection is to ensure that the information collected are of quality and reliable and can be subjected to statistical analysis.

4.5 SOURCES OF DATA

I. Primary Data

These are raw data which has just been collected from the source and has not gone in any kind of statistical treatment like sorting and tabulation. It may be referred to as firsthand information.

The important sources of primary data are primary units such as individuals, households, published sources (data that are available in print or in electronic including those found on internet web sites), basic experimental units.

Methods used to collect primary data can be any of the following:

1. **Survey** – most used to assess thoughts, opinions and feelings. It is often used in social sciences, management, marketing and psychology to some extent. Some ways to conduct a survey are as follows:

a. **Questionnaire** – Interviewing involves asking questions and getting answers from participants in a study. It can be face-to-face interviews and face-to-face group interviewing. Questions may consist of open-ended or close ended questions. It can be conducted via telephone, mail, personal appearance, e-mail or through fax.

b. **Interview** – face to face conversation with the respondent.

2. **Experiments** – most suitable for medicine, psychological studies, and nutrition.

3. **Observations** – can be done in natural settings as well as in artificially created environment.

II. Secondary Data

These are data which has already been collected, sorted, tabulated, and has undergone a statistical treatment and is available in either published or unpublished form. For instance, those data that can be found from the review of literature, from books, biographies, archives. They are called fabricated or tailored data.

The following are sources of secondary data:

- 1. Government Offices and Semi government offices
- 2. Teaching and research organizations, (NSO, PSA etc)
- 3. Books, Biographies, Research journals, articles, and newspapers
- 4. Internet and websites
- 5. Libraries, data bases, etc.

4.6 TWO BROAD CATEGORIES OF DATA

I. QUALITATIVE DATA, (Categorical)

Qualitative data are mostly non-numerical and usually descriptive or nominal in nature. They may be observed and described by words, pictures, and symbols but not numbers, because they cannot be expressed in numerical form and therefore cannot be calculated or computed. Often (not always) such data captures feelings, emotions, or subjective perceptions of something.

Example 2: Hair color, Gender, ethnic groups, nationality, educational qualification, intelligence, honesty, marital status and other attributes of the population.

II. QUANTITATIVE DATA, (Numerical)

Quantitative data is numerical in nature and can be mathematically computed. It answers key questions “how many”, “how much” and “how often”.

Quantitative data are easily amenable to statistical manipulation and can be represented by a wide variety of statistical types of graphs and charts such as line, bar graph, scatter plot, etc.

Example 3: Number of enrollees in your course this school year, height and weight of children, distance travelled from home to office, scores in an examination, prices of commodities, salaries of workers, etc.

Table 4.0: Types of data based on their mathematical properties

Based on their Mathematical Properties		Examples
TYPES OF DATA Ordered with their increasing ➤ Accuracy ➤ Powerfulness of measurement ➤ Preciseness ➤ Wide application of statistical techniques	NOMINAL	Gender (Male, Female)
		Marital Status (Single, Married, Widowed, Separated)
		Nationality (Filipino, Chinese, Korean)
	ORDINAL	Rank in competition (First, Second, Third)
		Rating (Excellent, Good, Fair, Poor)
		Economic Status (Low, Medium, High)
	INTERVAL	Temperature in degree Celsius
		Standardized Examination Scores
		Weighing Scales showing equal intervals
	RATIO	Weight of Toddlers
		Heights of growing children
		Age of human

Nominal Data

Nominal data represents discrete units and are used to label variables that have no quantitative value. It is used for *classification purposes*. Numbers or letters may be used to represent nominal measurements.

Nominal scales are said to be the *least powerful* in measurement with no arithmetic origin, order, direction, or distance relationship, reason why it has limited or restricted use.

Ordinal Data

This type of data represents discrete and ordered units. The order of the values is what is important and significant, but the differences between each one is not really known.

Ordinal scales are typically the measure of non-numeric concepts like satisfaction, happiness, discomfort, etc., and because they only show sequence, arithmetic cannot be done with it.

Interval (Score/Mark Data)

These are set of numerical measurements in which the distance between numbers are known, constant size or equal. But they do not have a “true zero”.

Interval scales are great, but ratios cannot be calculated. Interval data are *more powerful than ordinal scale* due to equality of intervals.

Ratio Data

Ratio data is defined as quantitative data, having the same properties as interval and definitive ratio between each data and absolute “zero” being treated as a point of origin, which means there can be no negative numerical value in ratio data.

The most precise data and allow for application of all statistical techniques.

*Summarizing the types of data: **Nominal** variables are used to “name” or label a series of values, **ordinal data** provide good information about order of choices, such as customer satisfaction survey. **Interval scales** give us the order of values and the ability to quantify the difference between each one. And **ratio scales** provide the order, interval values plus the ability to calculate ratio since a “true zero” can be defined.*

DISCRETE DATA AND CONTINUOUS DATA

Discrete Data

Discrete data has values that are distinct and separate. It represents items that *can be counted* and *only involves integers*. It should be converted to continuous data when possible to obtain a high level of information and details.

Example 4: The number of students in a Mathematics class, number of days in a year, number of players in a volleyball team, number of passengers in LRT 1 during rush hours, etc.

Continuous Data

Continuous data represents measurements in which values cannot be counted but they *can be measured*. It is more precise, more informative and can remove estimation and rounding of measurements, but it is often more time consuming to obtain.

Example: Heights and weights of students, number of seconds after the start of a race

Note: A good great rule for defining if a data is continuous or discrete is that if the point of measurement can be reduced in half and still make sense, the data is continuous.

4.7 FREQUENCY

The number of times a data point occurs in the set of data or the numerical count of data in each class interval.

Frequency Distribution

A table that list each data point and its frequency. It shows the frequency, or number of occurrences, in each of several categories. Frequency distributions are used to summarize large volumes of data values.

Relative Frequency

Relative frequency is the frequency of a data point expressed as a percentage of the total number of data points. It can be written as fractions, percent or decimals.

Cumulative Frequency

Cumulative frequency is the accumulation of the previous relative frequencies. To obtain the current frequencies just add all the previous relative frequencies.

Raw Data

Raw data is data that is not usually summarized or organized in any meaningful way. Often it is data as it is collected or recorded without any particular order except time of observation or sequence of observation.

Class interval

Class intervals are one way of categorizing raw data according to numerical constant intervals.

$$Class\ (Interval)width, W = \frac{highest\ value - lowest\ value}{number\ of\ classes} = \frac{Range}{number\ of\ classes}$$

Number of Classes

It is a common practice to keep the number of classes between 5 and 20 classes. However, in deciding the approximate number of classes, H.A. Sturges suggests the following formula:

$$K = 1 + 3.322 \log N$$

Where: K - represents the number of classes

log N – logarithm of the total number of classes

Example: If the total number of observations is 50, the number of classes would be

Solution: $K = 1 + 3.322 \log N$
 $K = 1 + 3.322 \log 50 = 6.644$

Therefore, the approximate number of classes is 7.

Example 5: Determine Frequency, relative frequency, frequency distribution of the following data:

1, 3, 6, 4, 5, 6, 3, 4, 6, 3, 6, 5, 5, 3

Solution: The Frequency, relative frequency, frequency distribution of the above data may be presented in Table 4.1, where x column represents the data.

Table 4.1: Frequency Table

x	Frequency, f	Relative Frequency, rf	Cumulative Relative Frequency
1	1	$\frac{1}{14} = 0.07$	0.07
3	4	$\frac{4}{14} = 0.29$	0.07+ 0.29 = 0.36
4	2	$\frac{2}{14} = 0.14$	0.36 + 0.14 = 0.5
5	3	$\frac{3}{14} = 0.21$	0.5 + 0.21 = 0.71
6	4	$\frac{4}{14} = 0.29$	0.71 + 0.29 = 1.00
	14		

4.8 CLASSIFICATION OF DATA

1. Ungrouped data

Ungrouped data are list of numbers that has not been organized into classes, or categories or into groups.

Example 6: Given the following ungrouped data, prepare the frequency distribution.

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

Solution: The given data can be arranged as shown in Table 4.2.

Table 4.2

No.	Frequency
0	4
1	9
2	14
3	8
4	9
5	2
	Total = 46

2. Grouped data

Grouped data are list of numbers that has been organized and arranged into classes and categories and some data analysis has been done.

Example 7: Frequency Distribution of Heights of Freshmen

Table 4.3: Heights of Freshmen

Height (cm)	Number of Students (Frequency)
140-150	78
151-160	172
161-170	39
171-180	11
	Total = 300

Example 7a: A record of 30 cars speed of a particular street are as follows with accuracy of 1 km/hr. Construct a frequency distribution for the given data.

62 58 58 52 48 53 54 63 69 63
57 56 46 48 53 56 57 59 58 53
52 56 57 52 52 53 54 58 61 63

(Source: http://www.uobabylon.edu.iq/eprints/publication_1_326_638.pdf)

Solution: The following are suggested steps for the construction of frequency distribution.

Step 1: Find the range

Highest value = 69

Lowest value = 46

Range = 69 – 46 = 23

Step 2: Find the number of class intervals using Sturges formula where N = 30

$K = 1 + 3.322 \log 30 = 1 + 3.322 (1.477) = 1 + 4.9066 = 5.9066$

Therefore, the number of classes = **6**

Step 3: Width of class interval, $W = \frac{23}{6} = 3.833 \approx 4$

Step 4: The class limits and all frequencies belong to each class interval maybe computed as follows:

First class: Lower limit = 46

Upper limit = $46 + 4 - 1 = 49$

Second class: Lower limit = $49 + 1 = 50$

Upper limit = $50 + 4 - 1 = 53$

Third Class: Lower limit = $53 + 1 = 54$

Upper limit = $54 + 4 - 1 = 57$

Step 5: Preparing the frequency table.

Class Interval	Frequency
46 - 49	3
50 - 53	8
54 - 57	8
58 - 61	6
62 - 65	4
66 - 69	1

Note: Tally sheet is omitted

4.9 METHODS OF DATA PRESENTATION

Table 4.4 Methods of Data Presentation

TEXTUAL METHOD	TABULAR METHOD	GRAPHICAL METHOD
<ul style="list-style-type: none">• Data are arranged from lowest to highest.• Stem-and-Leaf Plot	<ul style="list-style-type: none">• Frequency distribution• Relative Frequency distribution• Cumulative Frequency distribution• Contingency table	<ul style="list-style-type: none">• Bar Graph• Histogram• Frequency polygon• Pie Chart• Less than or greater than Ogive

I. Textual Presentation of Data

Data can be presented using paragraphs or sentences. It lists and identifies characteristics, significant figures, and important features of data.

Example 8: Scores in Quiz 1 of MMW class (40 students),

33	44	43	29	50	35	43	25
35	48	28	50	33	44	25	29
43	33	25	20	35	20	20	25
37	37	40	30	45	43	29	25
20	48	37	29	33	25	20	20

Solution: Rearranging the given scores from lowest to highest,

20	20	25	29	33	37	43	45
20	25	25	29	33	37	43	48
20	25	28	30	35	37	43	48
20	25	29	33	35	40	44	50
20	25	29	33	35	43	44	50

Its textual form can be expressed as follows:

In the MMW class of 40 students, 2 students got a perfect score, 34 students got a passing score of 25 points and above while 6 students got the lowest score of 20. Generally, 85% (34/40) of the total number of students passed Quiz 1.

Stem-and-leaf Plot

Stem-and-leaf plot is a table which sorts data according to a certain pattern. It consists of separating a number into two parts. For a two-digit number, the stem consists of the first digit and the leaf consists of the second digit. For three-digit number, the stem consists of the first two digits, and the leaf consists of the last digit. For a one-digit number, the stem is zero.

Example 9: Shown in Table 4.5 is a table of Steam-and-leaf of the previous problem.

Solution: First order the data, then separate the data by the first digit and last use the first digit as the stem and trailing digit as a leaf, as shown in Table 4.5

Table 4.5

Stem	Leaves
2	0,0,0,0,0,0,5,5,5,5,5,8,9,9,9,9
3	0,3,3,3,3,5,5,7,7,7
4	0,3,3,3,3,4,4,5,8,8
5	0,0

In the above Stem-and-leaf plot, there are 11 top scorer: 50, 50, 48, 48, 45, 44, 44, 43, 43, 43, 43 while the lowest 12 scores are 20, 20, 20, 20, 20, 20, 25, 25, 25, 25, 25, 25.

II. Tabular Presentation of Data

Tabulation can be in the form of simple tables or frequency distribution table.

1. Frequency Distribution Table

It is a table which shows data that are arranged into different categories or classes and their corresponding number of cases.

Frequency Distribution of Ungrouped Data

Example 10: Table 4.6 shows frequency distribution for the number of hours that 50 graduating high school students spend on social media sites per day

Table 4.6

Frequency Distribution Graduating High School Students	
No. of Hours	Frequency
1	6
2	15
3	22
4	4
5	3
	Total = 50

Frequency Distribution of Grouped Data

Table 4.7

Raw Scores, Quiz 1 in MMW Class		
Mark	Frequency	Cumulative Frequency
1 - 10	3	3
11 - 20	5	8
21 - 30	11	19
31 - 40	12	31
41 - 50	19	50
51 - 60	11	61
61 - 70	7	68
71 - 80	2	70

Figure 4.9

Contingency Table

A contingency table, sometimes called a two-way frequency table, is a tabular mechanism with at least two rows and two columns used in statistics to present categorical data in terms of frequency counts. The intersection of a row and a column of a contingency table is called a **cell**.

Example: The contingency table shown has two rows and five columns (not counting header rows/columns) and shows the results of a random sample of 2,200 adults classified by two variables, namely gender and favorite way to eat ice cream.

Gender	Cup	Cone	Sundae	Sandwich	Other
Male	592	300	204	24	80
Female	410	335	180	20	55

Source: <https://mathworld.wolfram.com/ContingencyTable.html>

One benefit of having data presented in a contingency table is that it allows one to perform basic probability calculations more easily, a feat made easier still by augmenting a summary row and column to the table as shown below:

Gender	Cup	Cone	Sundae	Sandwich	Other	Total
Male	592	300	204	24	80	1200
Female	410	335	180	20	55	1000
Total	1002	635	384	44	135	2200

4.10 GRAPHICAL METHOD

1. Bar Graph

The set of numbers (data) are illustrated using set of rectangular bars. It can be plotted vertically or horizontally, and it can be simple, multiple, or component type.

Example: Shown in figure 4.1 and figure 4.2 are Birth weights of newly born babies.

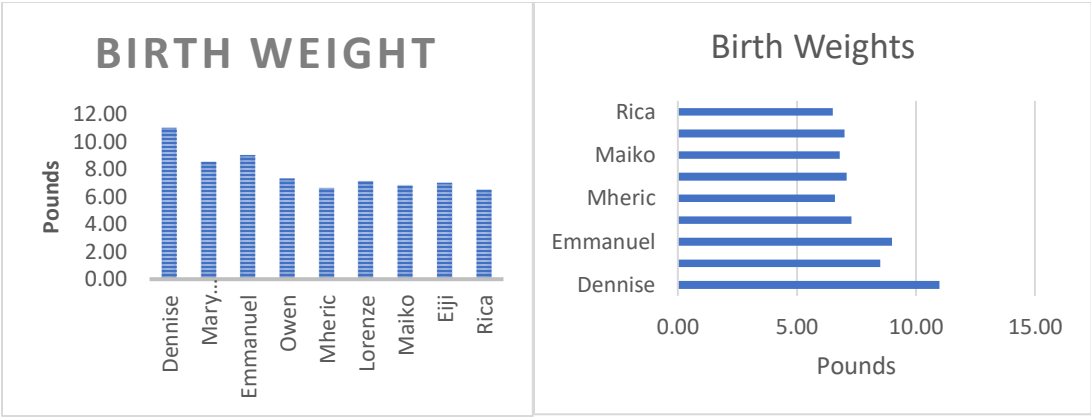


Figure 4.1

Figure 4.2

2. Histogram

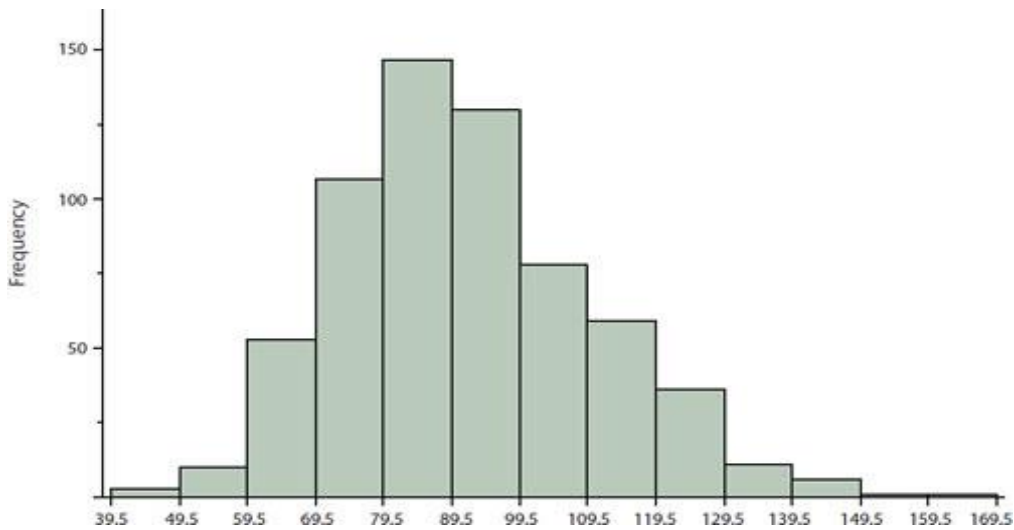
It is a pictorial diagram of frequency distribution. The class intervals are reflected along the x-axis, and the frequencies along the y-axis using series of bars. It looks like a bar chart but there are important differences between them.

Example: The Psychology test given to 642 students consist of 197 items, each graded as "correct" or "incorrect" with the students' scores ranged from 46 to 167. The test scores arranged in a simple frequency in Table 4.8. for the purpose of creating a Histogram.

Table 4.8 Grouped Frequency Distribution of Psychology Test Scores		
Lower Limit	Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

Table 4.8 shows the first interval is from 39.5 to 49.5, the second from 49.5 to 59.5, etc. There are three scores in the first interval, 10 in the second, etc. Class intervals of width 10 provide enough detail about the distribution to be revealing without making the graph too "choppy." Placing the limits of the class intervals midway between two numbers (e.g., 49.5) ensures that every score will fall in an interval rather than on the boundary between intervals.

Figure 4.3



Source: http://onlinestatbook.com/2/graphing_distributions/histograms.html

In the above histogram, the class frequencies are represented by bars and the height of each bar corresponds to its class frequency.

3. Cumulative Frequency Graph, (Ogive)

A cumulative frequency graph, also known as **Ogive**, is a curve showing the cumulative frequency for a given set of data. In Figure 4.4, the cumulative frequency is plotted along the y-axis and the data is plotted along the x-axis.

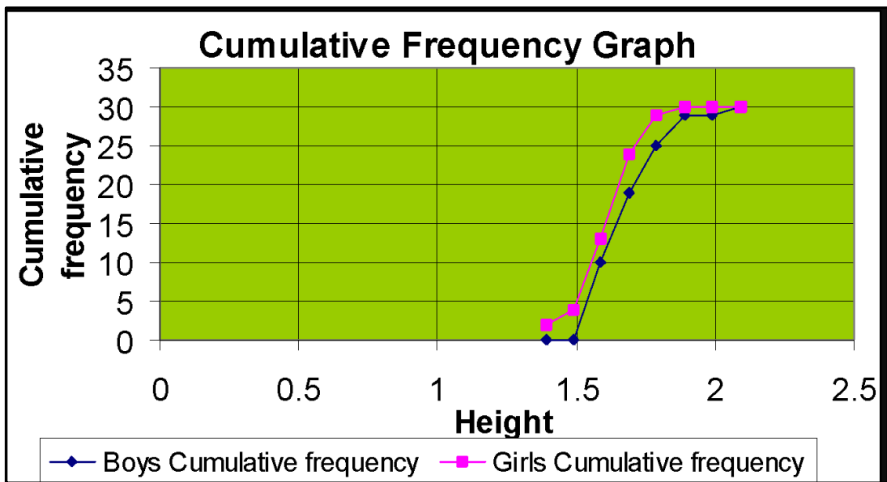


Figure 4.4

Source:http://static3.mbtfiles.co.uk/media/docs/newdocs/gcse/maths/data_handling/height_and_weight_of_pupils_and_other_mayfi_eld_high_school_investigations/34766/html/images/image01.png

4. Frequency polygon

It is a graph obtained by joining the mid-points of histogram blocks using line segments. Table 4.8 shows the scores in Quiz 1 of 3 sections in MMW and its equivalent frequency polygon in Figure 4.5

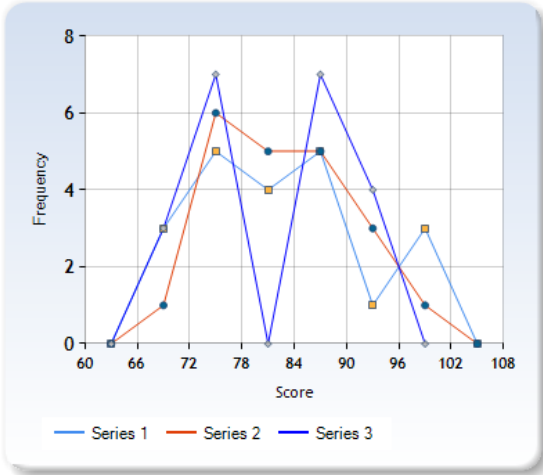


Figure 4.5

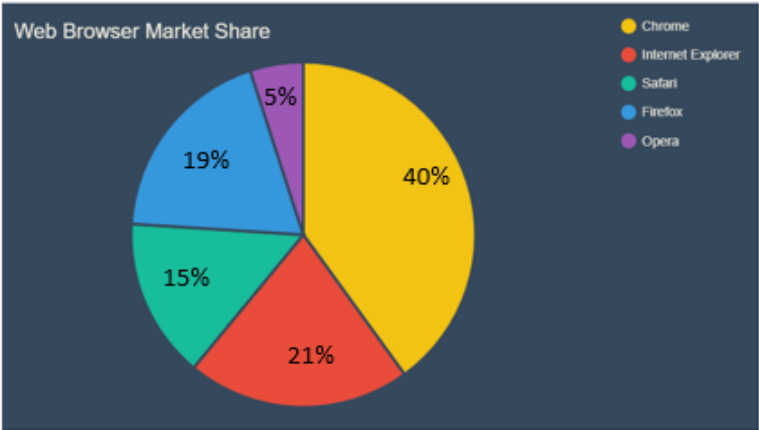
Table 2.8: SCORES IN QUIZ 1, MMW

Section 1	Section 2	Section 3
80	80	70
79	80	74
67	75	75
95	75	68
98	78	67
100	78	91
75	69	93
66	72	95
75	75	94
75	88	88
78	95	89
86	95	87
86	96	86
85	91	88
70	88	85
79	88	84
74	86	72
98	87	75
86	82	73
88	77	73
77	77	75

5. Pie Charts

It is a chart in circular form that uses “pie slices” to illustrate the relative sizes of data. The chart is divided into sectors of circle where each sector shows the relative size of each data. In figure 4.6 shows is an example of a pie chart





































Figure 4.6



Source:<https://www.chartblocks.com/en/support/faqs/faq/when-to-use-a-pie-chart>

6. **Pictogram / Pictograph** A chart that uses pictures to represent data. Pictograms are illustrated in the same way as bar charts, but instead of bars they use columns of pictures to show the numbers involved.

Example: The following pictograph shows the number of students using the different types of transportation to go to school.

Walking	
Tricycle	      
Bus	         
Private Car	    
LRT/MRT	            
 - represents 10 students	

- a.) How many students go to school by private car?
- b.) If the total number of students involved in the survey is 400, how many symbols must be drawn for the students walking to school?
- c.) What is the percentage of students who cycle to school?

Solution:









































- a.) $4.5 \times 10 = 45$ students go to school by private car.
- b.) Calculating for the total number of students using the different types of transportation,

Total = Walking + Tricycle + Bus + Private Car + LRT/MRT

$400 = \text{Walking} + (7 \times 10) + (9.5 \times 10) + (4.5 \times 10) + (13.5 \times 10)$

Walking = $400 - 70 - 95 - 45 - 135 = \mathbf{55 \text{ students}}$

Therefore, the number of symbols that must be drawn for students who walk to school is shown below:

Walking	     
Tricycle	     
Bus	         
Private Car	    
LRT/MRT	           
 - represents 10 students	

c.) The percentage of students who cycle to school = $\frac{70}{400} \times 100 = \mathbf{17.5\%}$

Example: Shown below is the employee's data of a certain company, where a pictograph was created with the aid of MS Excel.

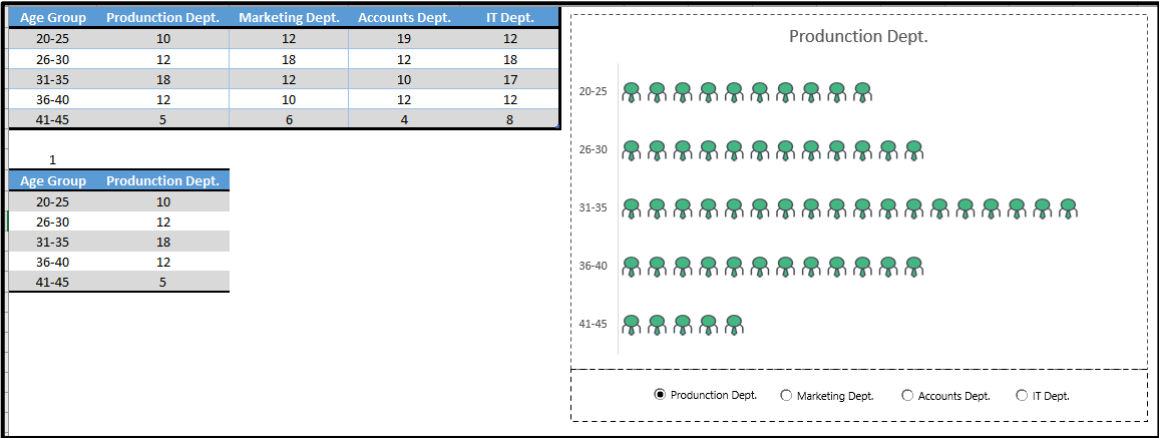


Figure 4.7

Source: <https://excelchamps.com/blog/how-to-make-a-pictograph-in-excel/>

7. Scatter Plot

It uses dots to represent values for two different variables. Scatter plots are used to observe relationships between variables and shows patterns when data are taken as a whole.

Example 11: What your car says about your salary?

Table 4.8

	X(Salary)	Y, (Car Price)
1	2,138,150	972,750
2	9,769,350	4,698,250
3	1,833,600	1,042,900
4	10,881,850	5,358,200
5	3,736,700	1,701,800
6	6,527,500	4,390,300
7	2,148,800	896,350
8	7,556,600	4,575,900
9	2,746,800	1,473,950
10	1,910,000	855,900
11	2,450,745	999,999
12	3,465,700	3,335,450
13	2,790,300	1,799,000
14	3,800,000	1,325,750

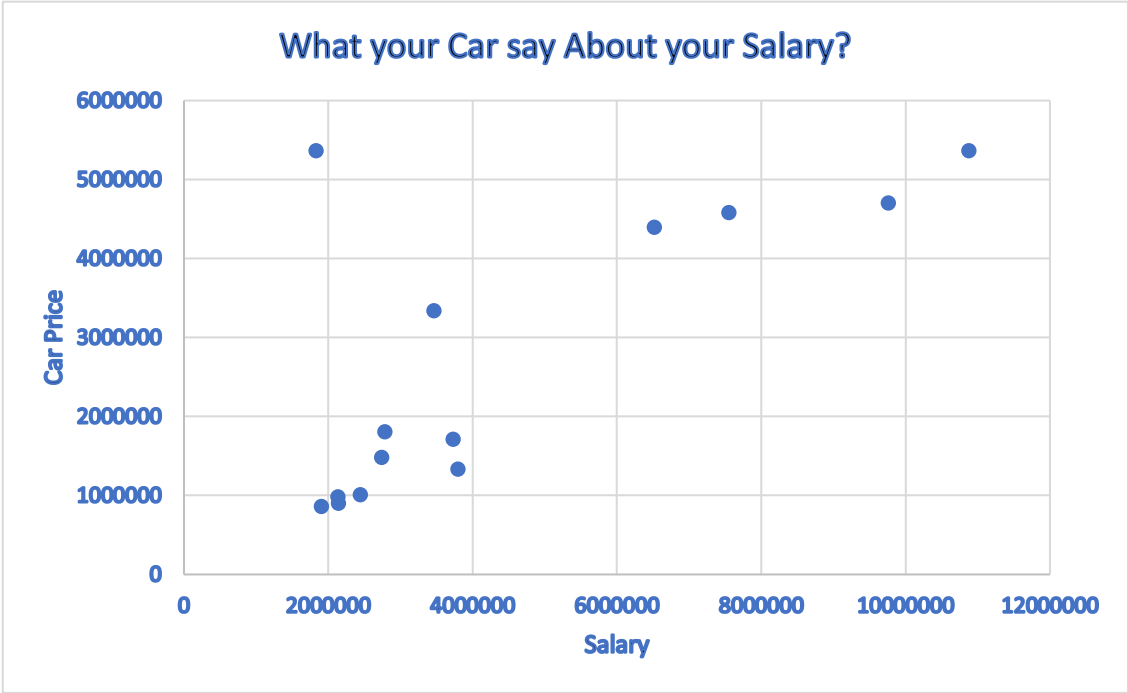


Figure 4.9

Review Exercises 4.0

Name _____ Score _____ Date _____

Course, Year & Section _____ Student no. _____ Professor _____

Solve the following problems as indicated.

1. The heights of 400 engineering students are given in the table 4.10, construct a frequency polygon.

Table 4.10

Heights, cm	Number of Students
140 - 150	75
150 - 160	156
160 - 170	128
170 - 180	41
Total	400

2. Table 4.11 represents the heights, in inches, of a sample of 100 male basketball players in NCR. Construct the cumulative frequency curve.

Table 4.11

Frequency Table of Soccer Player Height	
Heights, inches	Number of Students
59.95 – 61.95	3
61.95 – 63.95	9
63.95 – 65.95	10
65.95 – 67.95	33
67.95 – 69.95	15
69.95 – 71.95	10
71.95 – 73.95	5
73.95 – 75.95	15
Total	100

3. Twenty people were asked the number of kilometers they commute to work every day.

2 5 7 3 2 10 15 20 7 10

18 5 12 13 12 4 5 10 18 4

With the above data, the following table was produced:

Table 4.12

Frequency of Commuting Distances

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
2	2	4/19	0.2105
3	3	3/19	0.1579
4	1	1/19	0.2105
5	3	3/19	0.1579
7	2	2/19	0.2632
10	3	4/19	0.4737
12	2	2/19	0.7895
13	1	1/19	0.8421
15	1	1/19	0.8948
18	1	1/19	0.9474
20	1	1/19	1.000

Problem:

- a. Is the table correct? If it is not correct, what is wrong?
 - b. True or False: Three percent of the people surveyed commute 3 kilometers. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
 - c. What fraction of the people surveyed commute 5 or 7 kilometers?
 - d. What fraction of the people surveyed commute 12 kilometers or more? Less than 12 kilometers? Between 5 and 13 kilometers, (does not include 5 and 13 miles)?
4. Draw the Ogive graph for the following set of data:

2, 7, 16, 21, 31, 3, 8, 17, 21, 55, 3, 13, 22,
55, 4, 14, 19, 25, 57, 6, 15, 20, 29, 58, 18

5. An insurance company requested a police report about color of cars that encountered accident during first quarter of the year. The report is shown in the table below:

Color of Car	Red	Blue	White	Black	Gray	Brown
Frequency of Car Accident	26	55	43	38	41	25

Construct a bar graph for the above police report. Also create its pie-chart.

6. Consider the following data on annual wages of 20 Security Guards assigned to different establishments in Metro Manila Area.

Annual Wages of Security Guards in Hundred Thousand of Pesos

168 162 216 210 162 180 180 210 204 156
210 204 168 216 216 202 162 204 168 180

Determine Frequency, relative frequency, frequency distribution.

7. The temperature, °C in Baguio for the month of March 2020 is shown in Table 4.13

Table 4.13

BAGUIO TEMPERATURE, °C MARCH 2020						
SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
1	2	3	4	5	6	7
33°	33°	32°	31°	34°	33°	35°
8	9	10	11	12	13	14
35°	36°	34°	34°	33°	35°	35°
15	16	17	18	19	20	21
33°	31°	32°	34°	34°	35°	35°
22	23	24	25	26	27	28
36°	35°	34°	34°	34°	34°	36°
29	30	31				
36°	34°	34°				

Source: <https://www.accuweather.com/en/ph/baguio-city/262309/march-weather/262309?year=2020>

Prepare a stem-and-leaf plot table for the above data.

8. Consider the table 4.14 that displays a selection of variables from a study data set and answer the questions that follow:

Table 4.14

ID	AGE	GENDER	HEIGHT	Blood Group	LDL	Feeling Happy?	Number of Children	Smoke?	Social Class
1	25	F	1.62	B	150	Agree	0	No	I
2	35	F	1.58	O	123	Strongly Agree	1	Yes	II
3	44	M	1.35	A	178	Disagree	3	Yes	I
4	28	F	1.54	AB	205	Disagree	0	No	III
5	3	M	1.35	O	229	Indifferent	2	Yes	I
6	42	M	1.21	B	215	Agree	2	Yes	IV
7	36	F	1.76	A	130	Strongly Disagree	1	No	IV
8	38	M	1.57	A	175	Disagree	1	Yes	V
9	30	M	1.47	AB	240	Indifferent	0	No	III
10	40	F	1.18	B	167	Strongly agree	6	No	I

*LDL – Low Density Lipoprotein

Source: <https://www.slideshare.net/WinonaEselBernardo/presentation-of-data-10958540>

Which of the above variable(s) are classified as quantitative variable(s)?

- a. ID

b. Age
- c. Gender

d. Height
- e. Blood group

f. LDL
- g. Feeling Happy?

h. Number of children
- i. Smoke?

J. Social Class
- k. None of the above

9. Consider Table 2.14, which variable(s) are classified as qualitative variable(s)?

- a. ID

b. Age
- c. Gender

d. Height
- e. Blood group

f. LDL
- g. Feeling Happy?

h. Number of children
- i. Smoke?

J. Social Class
- k. None of the above

10. Which of the above variable(s) are classified as ordinal variable(s)?

- a. ID

b. Age
- c. Gender

d. Height
- e. Blood group

f. LDL
- g. Feeling Happy?

h. Number of children
- i. Smoke?

J. Social Class
- k. None of the above

4.11 DESCRIPTIVE STATISTICS

MEASURES OF CENTRAL TENDENCY

The measures of central tendency are ways of describing the central position of a set of data. They are sometimes called measures of central location or summary statistics.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

I. Mean

It is the sum of all measurements divided by the total number of measurements in the given data. It is the most used measure of central position, the point on which a distribution would be balance.

Mean for Ungrouped Data, \bar{x}

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

The above formula is also called *arithmetic mean, simple mean or average*.

- Where, \bar{x} – the mean of the measurements, x
- $\sum x$ – the summation of all measurements
- n – the total number of measurements

Example 11: The scores in a short quiz of 30 students in Mathematics are as follows, find the mean score of the class.

37, 33, 33, 32, 29, 28, 23, 22, 22, 22, 21, 21, 21, 20, 20,
19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6

Solution: Substitute the given scores to the formula $\bar{x} = \frac{\sum x}{n}$, where n = 30

$$\bar{x} = \frac{37 + 33 + 33 + 32 + 29 + 28 + 23 + 22 + 22 + 22 + 21 + 21 + 21 \dots + 6}{30}$$
$$\bar{x} = \frac{606}{30} = 20.2$$

Therefore, the mean score of the class, $\bar{x} = 20.2$

The Weighted Mean

Weighted mean is a mean that is calculated with extra weight given to the sample data.

$$\bar{x} = \frac{\sum wx}{\sum w}$$

Where w – the weight of the given data

Example 12: The numbers 1, 2, 3, 4 and 5 have weights 0.1, 0.25, 0.3, 0.15 and 0.2, respectively. Calculate its weighted mean?

Solution: Substitute the given numbers with their corresponding weights as follows:

$$\bar{x} = \frac{\sum wx}{\sum w} = \frac{1(0.1) + 2(0.25) + 3(0.3) + 4(0.15) + 5(0.2)}{0.1 + 0.25 + 0.3 + 0.15 + 0.2}$$
$$\bar{x} = \frac{0.1 + 0.5 + 0.9 + 0.6 + 1}{1} = 3.1$$

Therefore, the weighted mean, $\bar{x} = 3.1$

Mean for Grouped Data

The mean for grouped data can be solved by any of the following formula:

METHOD'S NAME	GROUPED DATA
Direct Method	$\bar{x} = \frac{\sum fx}{n}$
Step-Deviation Method	$\bar{x} = A + \frac{\sum fu}{\sum f}(h)$

Where: \bar{x} – mean of grouped data
 x – class mark or the midpoint of the upper and the lower limit of the class
 n – total number of observations equals to $\sum f$
 f – indicates frequency of different groups/class
 A – indicates assumed mean
 $u = \frac{x - A}{h}$
 u – indicates step-deviation/unit code deviation
 h – indicates size of class
 $\sum fx$ – summation of the product of the class marks and corresponding frequencies.

Example 13: The following data shows the number of days that students were absent from school due to sickness in the previous school year. Use the direct method to compute the mean.

No. of days absent due to sickness	1 - 5	6 - 10	11 - 15	16 – 20	21 - 25
Frequency	12	11	10	4	3

Solution: First, prepare the table as follows,

No. of days absent due to sickness	F	midpoint, x	fx
1 - 5	12	3	36
6 – 10	11	8	88
11 – 15	10	13	130
16 – 20	4	18	72
21 - 25	3	23	69
	n = 40		$\sum fx = 395$

$$\bar{x} = \frac{\sum fx}{n} = \frac{395}{40} = 9.875$$

Therefore, the mean number of student’s absences due to illness is **9.875**.

Example 14: The following data shows the distance traveled by 100 commuters from house to work.

Distance in kilometers	1 - 10	11 - 20	21 - 30	31 - 40	41 - 50
No. of Commuters	10	17	42	18	13

Calculate the arithmetic mean by step-deviation method.

Solution: The given data can be tabulated as follow, where the assumed mean, $A = \frac{21+30}{2} = 25.5$

Distance, km	Number of Commuters f	Mid points x	$u = \frac{x - A}{10}$	fu
1 – 10	10	5.5	-2	-20
11 – 20	17	15.5	-1	-17
21 – 30	42	25.5	0	0
31 – 40	18	35.5	1	18
41 - 50	13	45.5	2	26
Total	$\sum f = 100$			$\sum fu = 7$

Solving for the arithmetic mean using the formula, $\bar{x} = A + \frac{\sum fu}{\sum f} (h)$

$$\bar{x} = 25.5 + \frac{7}{100} (10) = 26.2$$

Note: The mid-point is a multiple of 5 and the difference from mid-point to mid-point is 10, it is the class size.

Example 15: Shown below is a table of scores of 50 students in MMW class. Find the arithmetic mean by a.) direct method b.) step-deviation method.

Scores	20 - 29	30 - 39	40 - 49	50 - 59	60 - 69	70 - 79	80 - 89
Frequency	1	5	12	15	9	6	2

Solution: The frequency table can be prepared as follows:

			Direct Method	Step-Deviation Method	
Marks	f	x	fx	$u = \frac{x - A}{10}$	fu
20 - 29	1	24.5	24.5	-3	-3
30 - 39	5	34.5	172.5	-2	-10
40 - 49	12	44.5	534	-1	-12
50 - 59	15	54.5	817.5	0	0
60 - 69	9	64.5	580.5	1	9
70 - 79	6	74.5	447	2	12
80 - 89	2	84.5	169	3	6
Total	50		2745		2

Where $A = \frac{50+59}{2} = 54.5$ $h = 10$

a.) Direct Method

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{2745}{50} = \mathbf{54.9}$$

b.) Step-Deviation Method

$$\bar{x} = A + \frac{\sum fu}{\sum f} \times h$$

Where $A = \frac{50+59}{2} = 54.5$ and $h = 10$,

$$\bar{x} = 54.5 + \frac{2}{50} \times 10 = \mathbf{54.9}$$

Advantages of Arithmetic mean

Arithmetic mean is based on observations and easy to calculate and determined for almost every kind of data. It is only slightly affected by the unstable values of samples.

Disadvantages of Arithmetic mean

It is highly affected by utmost values. It is not the proper way of expressing averages into ratios and percentages and for highly skewed distributions is not a proper way to compute its average.

II. MEDIAN for UNGROUPED DATA

Median is defined as the middle value of the data when the data is arranged in ascending or descending order. When the number of observations is odd, the middle number is the median, but when the number of observations is even, the median is the average of the two middle numbers.

The formula is as follows:

If the number of observations is odd: $Md = \text{value of } \left(\frac{n+1}{2}\right) \text{th data}$

If the number of observations is even: $Md = \frac{\left(\frac{n}{2}\right)th + \left(\frac{n}{2}+1\right)th}{2}$

Where Md – is the median of ungrouped data

n – is the total number of observations

Example 16: Find the median of the following data.

37, 25, 29, 43, 21, 17, 35

Solution: Arrange the data in ascending order.

17, 21, 25, **29**, 35, 37, 43

There are 7 terms in the given, which is odd, **therefore the median is the middle (4th) term which is 29 or Md = 29.**

Note: The number of terms having values greater than or equal 29 is the same as the number of terms having values less than or equal to it.

Example 17: Given the following data, find the median.

45, 55, 35, 65, 85, 75

Solution: Arrange the data in ascending order.

35, 45, **55, 65**, 75, 85

The number of terms is 6 which is even, which means the 3rd and 4th are middle terms. Median is the average value of these terms,

$$Median = \frac{55 + 65}{2} = \frac{120}{2} = 60$$

Hence, Md = 60.

Note: The number of terms having values greater than or equal to 60 is the same as the number of terms having values less than or equal to it.

MEDIAN OF GROUPED DATA

The formula for Median of grouped data is as follows:

$$Md = L + \frac{i\left(\frac{n}{2} - f <\right)}{f}$$

Where: L – lower limit of median class

i - class size of the distribution

n – number of observations

$f <$ - less than cumulative frequency of the class below the median class

f – frequency of the median class

Example 18: Given the following frequency distribution, find the median.

Raw Score	f
75 -79	3
70 - 74	1
65 - 69	7
60 -64	9
55 - 59	6
50 - 54	4

Solution: Add a third column for the less than cumulative frequency to the above table. The third column is obtained by adding the frequencies of the class starting from the lowest class up to the topmost class. The cumulative frequency of the topmost class must be equal to the number of observations

Raw Score	f	f <
75 -79	3	30
70 - 74	1	27
65 - 69	7	26
60 -64	9	19
55 - 59	6	10
50 - 54	4	4
	n = 30	

For the median class, consider the class opposite the third column ($f <$) which is immediately greater than $\frac{n}{2}$. Since $\frac{n}{2} = 15$, the entry in the third column that is immediately greater than it is 19, and therefore the median class is 60 – 64. The lower limit of the median class is 60 and the exact lower limit of the median class is $60 - 0.5 = 59.5$. The class size $i = 5$ and $f < = 15$, then Md is as follows,

$$Md = L + \frac{i\left(\frac{n}{2} - f <\right)}{f} = 59.5 + \frac{5(15 - 10)}{9}$$

Hence, **Md = 62.27**

MODE OF UNGROUPED DATA

The mode is the number that appears most frequently in a data set. A set of data may have one mode, (unimodal) two modes (bimodal), 3 modes (trimodal) or more (multimodal) or no mode at all (zero mode).It is denoted by the symbol M_o and it can be obtained by inspection of the given set of data.

Example 19: Given the following sets of numbers, determine its mode.

- a. 2, 3, 5, 9, 15, 15, 15, 26, 27, 39, 42
- b. 80, 79, 91, 80, 85, 76, 79, 74, 91
- c. 5, 10, 15, 20, 25, 30, 30, 25, 20, 15, 10, 5

Solution:

- a. The number that occur most frequently is 15, therefore the **M_o = 15**.

2, 3, 5, 9, **15, 15, 15**, 26, 27, 39, 42

- b. There are three numbers that occur most frequently, 79, 80 and 91 therefore the **M_o = 79, 80, 91**. The mode of the given set of data is trimodal.

74, 76, 79, 79, 80, 80, 85, 91, 91, 93

- c. There is no mode in the following set of data, **M_o = 0**

5, 10, 15, 20, 25, 30, 30, 25, 20, 15, 10, 5

MODE OF GROUPED DATA

The mode of grouped data can be computed using the following formula:

$$M_o = L_B + i \left(\frac{d_1}{d_1 + d_2} \right)$$

Note: This is a formula for Moments of Force Method.

Where M_o – mode of grouped data


L_B – lower boundary of the modal class (also called exact lower limit of modal class)

i – class size

d₁ – difference between frequency of the model class and the class below it.

d₂ – difference between frequency of the modal class and the class above it.

Example 20: A record of Elite Coffee Shop shows the number of orders of Brewed Coffee made by its customers per hour. Find the mode.

Number of Brewed Coffee	Frequency	 <div>Image Source: https://youronevoicecanmakeadifference.files.wordpress.com/2012/06/cup-of-fresh-brewed-coffee-photo-credit-fanpop1.jpg</div>
0 - 2	2	
3 - 5	3	
6 - 8	6	
9 - 11	7	
12 - 14	5	
15 - 17	3	
18 - 20	2	

Solution: Locate the modal class, as shown below the modal class is 9 – 11, having the highest frequency of 7,

Number of Brewed Coffee	Frequency
0 - 2	2
3 - 5	3
6 - 8	6
9 - 11	7
12 - 14	5
15 - 17	3
18 - 20	2

And the L_B , lower boundary of the modal class = $9 - 0.5 = 8.5$

$$d_1 = 7 - 6 = 1$$

$$d_2 = 7 - 5 = 2$$

$$i = 3$$

$$M_o = L_B + i \left(\frac{d_1}{d_1 + d_2} \right) = 8.5 + 3 \left(\frac{1}{2 + 1} \right) = 9.5$$

Advantages of Mode

Mode is easy to understand and calculate and is useful for qualitative data. It is not affected by extreme values. It is easy to identify in a data set and in a frequency distribution. It can be computed in an open -ended frequency table and can be located in a graph.

Disadvantages of Mode

Mode is defined when there are no values repeated in a data set. It is not based on all values. It unstable when data consist of a small number of values.

Example 21: Find the arithmetic mean, median and mode of the following set of data:

73, 78, 73, 74, 73, 76, 74, 81

Solution: Solving first the arithmetic mean,

$$\bar{x} = \frac{\sum x}{n} = \frac{73 + 78 + 73 + 74 + 74 + 73 + 76 + 81}{7} = 75.25$$

To solve for the median, arrange the given data in ascending order, there are 8 items in the data which means the median is the average of the value 4th and 5th item in the sequence,

73, 73, 73, **74, 74**, 76, 78, 81

$$Md = \frac{74 + 74}{2} = 74$$

By inspection of the given data, 73 occurs most frequently, therefore **M_o = 73**

Note: In many cases, the modal value will differ from the average value in the data.

Exercises 4.1

Name _____ Score _____ Date _____
Course, Year & Section _____ Student no. _____ Professor _____

Solve the following problems.

1. Find the mean of the following data in the frequency table.

x	f
10	6
23	13
27	15
30	17
35	12

2. Shown in the table below is the Age distribution of newly hired health workers to be deployed to different hospitals in USA. Use the Step-Deviation Method to compute for mean deployment rate of health workers.

Age Group	Number of Health workers
25 – 29	18
30 – 34	21
35 – 39	35
40 – 44	40
45 – 49	36
50 – 54	29
55 – 59	15

3. The final grade of Emmanuel in every subject is indicated in his report card shown below, use weighted mean to calculate for Emmanuel's GPA.

Subjects	No. of Units	Final Grade
Oral Communication in Context	3	88
Komunikasyon at Pananaliksik sa Wika at Kulturang Pilipino	3	90
General Mathematics	3	85
Earth Science	3	89
Introduction to the Philosophy of the Human Person	3	92
Empowerment and Technologies	3	91
Pre-Calculus	3	87

Note: P.E. is not included in the computation of GPA.

4. Find the median of the following given data:

- a. 13, 14, 15,1 3, 16, 17, 12.
- b. 42, 44, 45, 43, 48, 47, 44, 40.

5. During the 4 months community quarantine due to pandemic the Popular Electric company was not able to do actual meter reading of electric consumption instead charged their customers based on average household consumptions for the last 4 months. If a household electric consumption for the last 4 months were 95kwh, 106.2 kwh, 98.9kwh and 101.3kwh. Find the number of kilowatt hour that the Popular electric company charged to the customer on the 5th month of pandemic.

6. The typing speeds of 25 secretaries in a big company are recorded below (in words per minute). Find the mode.

35, 43, 39, 46, 43, 47, 38, 51, 43, 38, 40, 45

7. The manager of a Burger House recorded the number of burgers sold per day in 2 weeks (below). Which of the following statements is true?
- 132, 121, 119, 116, 130, 121, 131, 117, 119, 135, 121, 129, 119, 134
- a. There is no mode. b. The mode is 119
c. The mode is 131 d. The modes are 119 and 121
8. Medric’s cumulative GPA for 3 semesters was 2.5 for 42 course units. His fourth semester was 3.0 for 15 course units. What is his cumulative GPA for all 4 semesters?
9. A Mathematics professor evaluates students on 5 quizzes, a project and final examination. Each quiz is equivalent to 10% of course grade, the project is 20% and the final examination is 30% of the course grade. A student test scores are 75, 82, 77, 88 and 85, project score is 95 and final examination of 82. Use weighted mean to compute for the student’s course average.
10. Find the mean, median and all the modes for the following frequency distribution.

Points scored in basketball game	Frequency
3	5
4	6
5	5
8	3
11	2
15	1
20	1

11. Given the frequency distribution of scores in the final examination of 50 engineering students in Calculus I. Compute the modal score in the final examination in Calculus 1.

Final Examination Scores	Number of Students
96 – 100	1
91 – 95	3
86 – 90	11
81 – 85	16
76 – 80	9
71 – 75	7
66 – 70	3

12. Organizers of the 5 kilometer Race recorded the participants running pace in minutes. Find the mean, median and all the modes of the runners.

Time, minutes	Number of Participants
80 - 89	10
70 - 79	13
60 - 69	21
50 - 59	26
40 - 49	19
30 - 39	7
20 - 29	4

13. The following data represent a sample of 10 scores on Statistics quiz:

16, 16, 16, 16, 16, 18, 18, 20, 20, 20

After the mean, median, mode, range and variance were calculated for the scores, it was discovered that one of the scores of 20 should have been an 18. Which of the following will change when the calculations are redone using the correct scores?

- a. Mean and range
- b. Median
- c. Mean and Variance
- d. Variance and range

4.12 MEASURES OF DISPERSION

The measure of dispersion provides information on how scatter or spread are the given set of data. It gives a clear idea about the distribution of the data. It shows how much of the data vary from their average value. The measure of dispersion shows if the distribution of data is homogeneous or heterogenous. It is also known as the measure of variability.

Some ways to measure dispersion are calculating the range, mean absolute deviation, variance, and standard deviation.

The range and quartile deviation express the scattering of observation in terms of distances while mean deviation and standard deviation expresses the variations in terms of the average of deviations of observations.

4.13 RANGE OF UNGROUPED DATA

Range is the simplest measure of dispersion. It is the difference between two extreme observations of the data set.

$$R = H - L$$

Where R – range

H – highest value in the set of data

L – lowest value in the set of data

Example 22: Compute range of the of the Final grades of 10 Male and 10 Female Grade 11 students in Statistics. Describe the result.

Male: 91, 98, 87, 76, 75, 82, 84, 78, 86, 83

Female: 92, 88, 83, 85, 82, 77, 91, 78, 82, 79

Solution: The highest final grade of Male students is 98 and the lowest grade is 75, while for female the highest final grade is 92 and the lowest grade is 73, therefore the range are,

Male: $R = 98 - 75 = 23$

Female: $R = 92 - 77 = 15$

From the above computation the range of male is 23 while that of female is 15, this means that the final grades of male students are more scattered while the final grades of female students are closer. This also indicates that the final grades of female students are more homogeneous than the final grades of male students.

RANGE OF GROUPED DATA

There are 2 ways to calculate the range of grouped data.

1. Range = Upper boundary of the highest class - Lower boundary of the lowest class

$$R = U_H - L_L$$

2. Range = midpoint of the highest class - midpoint the lowest class.

$$R = M_h - M_L$$

Example 23: Given the weights of Female students in a PE class, find its range and coefficient of range.

Weight (Pounds)	100 - 104	105 - 109	110 - 114	115 – 119	120 - 124	125 - 129
Number of Students	8	13	19	43	10	7

Solution: Preparing the table as follows:

Weight (lbs)	Class Boundaries	Mid Value	No. of Students
100 – 104	99.5 – 104.5	102	8
105 – 109	104.5 – 109.5	107	13
110 – 114	109.5 – 114.5	112	19
115 – 119	114.5 – 119.5	117	43
120 – 124	119.5 – 124.5	122	10
125 - 129	124.5 – 129.5	127	7

Method 1: $R = U_H - L_L = 129.5 - 99.5 = \mathbf{30\ lbs.}$

Method 2: $R = M_H - M_L = 127 - 102 = \mathbf{25\ lbs.}$

Note: Range is the simplest method and easily understood measure of dispersion, but it is a poor measure of dispersion and does not give a good picture of the overall spread of the observations with respect to the center of the observations.

MEAN ABSOLUTE DEVIATION

Mean deviation is the arithmetic mean of the absolute deviations of the observations from a measure of central tendency.

$$MAD = \frac{\sum |x - \bar{x}|}{n}$$

Where MAD – Mean absolute deviation
 X – given value in the set of ungrouped data
 \bar{x} – mean of ungrouped data
 n – total number of measurements
 $\sum |x - \bar{x}|$ – summation of the absolute deviation of the values from the mean.

Example 23: The record of the 10 fastest participants of 5 Kilometer Fun Run in Manila are as follows 50, 60, 75, 80, 90, 105, 125, 145, 165, 200 all in minutes. Find the mean absolute deviation.

Solution: We first solve for \bar{x} ,

$$\bar{x} = \frac{\sum x}{n} = \frac{50+60+75+80+90+105+125+145+165+200}{10} = \frac{1095}{10} = \mathbf{109.5}$$

Now, solving for other items in the formula and presenting in tabulated form,

X	$x - \bar{x}$	$ x - \bar{x} $
50	-59.5	59.5
60	-49.5	49.5
75	-34.5	34.5
80	-29.5	29.5
90	-19.5	19.5
105	-4.5	4.5
125	15.5	15.5
145	35.5	35.5
165	55.5	55.5
200	90.5	90.5
n = 10		$\sum x - \bar{x} = 394$

Substituting to the MAD formula,

$$MAD = \frac{\sum |x - \bar{x}|}{n} = \frac{394}{10} = 39.4$$

The mean absolute deviation is **39.4**

4.14 MEAN ABSOLUTE DEVIATION OF GROUPED DATA

The formula for mean absolute deviation is as follows,

$$MAD = \frac{\sum f|x - \bar{x}|}{n}$$

Where MAD – mean absolute deviation for grouped data

x – class mark

\bar{x} – mean of grouped data

n – total number of measurements

f – frequency of the class

$\sum f|x - \bar{x}|$ – summation of the product of frequency and absolute deviation of the class marks from the mean

Note: The class mark of the class is the average of the lower and upper boundary of the class.

Example 24: Consider the previous problem, where the weights of Female students in a PE class are given, find its MAD.

Weight (Pounds)	100 - 104	105 - 109	110 - 114	115 – 119	120 - 124	125 - 129
Number of Students	8	13	19	43	10	7

Solution: We first solve for \bar{x} using the short cut method of mean of grouped data,

Weight, (lbs)	F	x	$u = \frac{x - A}{5}$	fu	$x - \bar{x}$	$f x - \bar{x} $
100 – 104	8	102	-3	-24	-14.55	116.4
105 – 109	13	107	-2	-26	-9.55	124.15
110 – 114	19	112	-1	-19	-4.55	86.45
115 – 119	43	117	0	0	0.45	19.35
120 – 124	10	122	1	10	5.45	54.5
125 - 129	7	127	2	14	10.45	73.15
	$\sum f = 100$			-45		474

$$A = \frac{115 + 119}{2} = 117$$

$$\bar{x} = A + \frac{\sum fu}{\sum f} = 117 + \frac{(-45)}{100} = \mathbf{116.55}$$

Substitute \bar{x} to MAD,

$$MAD = \frac{\sum f|x - \bar{x}|}{n} = \frac{474}{100} = 4.74$$

Note: The merit of mean deviation is based on all observations, it provides minimum value when the deviations are taken from the median and independent of change of origin. Its demerits, it is not easily understandable, its calculation is time consuming, negative sign is useless for mathematical treatment.

4.15 VARIANCE and STANDARD DEVIATION

Variance is another measure of dispersion. It is equal to the square of the standard deviation of the given set of data.

Standard deviation is defined as the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean.

It is also referred to as root mean square deviation.

VARIANCE AND STANDARD DEVIATION FOR UNGROUPED DATA

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} \text{ and } s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Where s^2 - variance of ungrouped data

S – standard deviation of ungrouped data

$x - \bar{x}$ – deviation of x values from the mean

n – number of measurements

$\sum (x - \bar{x})^2$ – summation of the squares of the deviations.

Example 25: The following numbers were obtained from the daily sales of pair of trendy school shoes for two weeks in one of its stores.

Week	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1	15	18	16	11	19	21	16
2	10	11	12	11	15	24	14

Calculate the variance of the above data.

Solution: We first solve for \bar{x} ,

$$\bar{x} = \frac{15 + 18 + 16 + 11 + 19 + 21 + 16 + 10 + 11 + 12 + 11 + 15 + 24 + 14}{14} = \frac{213}{14}$$

$\bar{x} = \mathbf{15.21}$

x	$x - \bar{x}$	$(x - \bar{x})^2$
10	-5.21	27.1
11	-4.21	17.7
11	-4.21	17.7
11	-4.21	17.7
12	-3.21	10.3
14	-1.21	1.5
15	-0.21	0.0
15	-0.21	0.0
16	0.79	0.6
16	0.79	0.6
18	2.79	7.8
19	3.79	14.4
21	5.79	33.5
24	8.79	77.3
n = 14		226.4

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{226.4}{14 - 1} = 17.415$$

Therefore, the variance of daily sales of trendy shoes is **17.415** and its standard deviation is **4.17**.

VARIANCE AND STANDARD DEVIATION FOR GROUPED DATA

The formula for variance and standard deviation for grouped data is shown below:

$$s^2 = \frac{\sum f(x-\bar{x})^2}{n-1} \quad \text{and} \quad s = \sqrt{\frac{\sum f(x-\bar{x})^2}{n-1}}$$

- Where s^2 – variance of grouped data
s - standard deviation of grouped data
f - frequency of the class
x - class mark
 \bar{x} – mean of the grouped data
n – total number of measurements
 $x - \bar{x}$ – deviation of the class mark from the mean of grouped data
 $\sum f (x - \bar{x})^2$ – summation of the product of the deviation and corresponding frequency of the class

Example 26: Consider the previous problem, the weights of Female students in a PE class are given, calculate it variance and standard deviation.

Solution: Calculating for \bar{x} and tabulating other computations as follows:

$$\bar{x} = \frac{\sum fx}{n} = \frac{11,475}{100} = 114.75$$

Weight, (lbs)	f	x	fx	$x - \bar{x}$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
100 – 104	8	102	816	-12.75	162.56	1300.5
105 – 109	13	107	1391	-7.75	60.06	780.8125
110 – 114	19	112	2128	-2.75	7.56	143.6875
115 – 119	43	117	5031	2.25	5.06	217.6875
120 – 124	10	122	1220	7.25	52.56	525.625
125 - 129	7	127	889	12.25	150.06	1050.438
	$\sum f = 100$		11,475			8,037.5

$$s^2 = \frac{\sum f(x-\bar{x})^2}{n-1} = \frac{8037.50}{100-1} = 81.19 \quad \text{then} \quad s = \sqrt{\frac{\sum f(x-\bar{x})^2}{n-1}} = \sqrt{\frac{8037.50}{99}} = 9.01$$

Review Exercises 4.2

Name _____ Score _____ Date _____

Course, Year & Section _____ Student no. _____ Professor _____

Solve the following:

1. Find the standard deviation of the sample of the following numbers obtained by sampling method.

a. 3, 5, 8, 13, 16, 20 b. 114, 121, 105, 129, 116, 126, 100, 98
2. The average number of hours spent by 20 neighbors in an exclusive subdivision watching television last Sunday were as follows: Find the range.

2, 4, 7, 5, 12, 9, 7, 8, 2, 2, 11, 10, 7, 7, 4, 4, 2, 1, 2, 1
3. The table below shows the wages of Employees of Company A and B.

	Company A	Company B
Number of Employees	950	1,100
Average Daily Wage	₱ 680.00	₱ 750.00
Variance in the distribution of Wages	100	145

a. Which company has a larger wage bill?
b. Calculate the average daily wage and the variance of the distribution of wages of all the employees in the firms A and B taken together.
4. Some 15 Students in Chemistry were observed to spend more time doing their experiments beyond their laboratory class time. The amount of time (minutes) spent by these students are as follows:

15, 28, 25, 48, 22, 43, 49, 34, 22, 33, 27, 25, 22, 20, 39

a. Find the Range, Standard Deviation, and Variance for the above data.

b. What does this information tell you about the variability of student's length of time doing their experiments beyond their laboratory class time? Is it homogeneous or heterogeneous?
5. The final scores in 10 Pin Bowling obtained by 6 players are given below:

Players	Score
Lorenze	81
Maiko	130
Eiji	80
Rica	109
Dennise	150
Owen	86

a. Find the Range, Standard Deviation, and Variance for the above data.

b. What does this information tell you about the variability of final scores in 10 Pin Bowling of the players? Is it homogeneous or heterogeneous?
6. The table below shows the total number of man-days lost to sickness during one week's operation of a small chemical plant.

Days Lost	1-3	4-6	7-9	10-12	13-15
Frequency	8	7	10	9	6

Calculate the variance and standard deviation of the number of lost days.

7. Use Step-deviation method to calculate the mean of the following data:

Quiz 1 Raw Scores	11 - 20	21 - 30	31 - 40	41 - 50	51 - 60	61 - 70	71 - 80
Number of Students	4	7	15	18	12	6	1

8. The variance of a sample of 121 observations equals 441. Find its standard deviation.

9. In a farm, 50 hogs are raised. The caretaker observed the weight gain of pigs who are fed with starter ration from weaning until two months of age are recorded in the table below:

Weight Gain, Kg	Number of Pigs
3.0 – 3.9	1
4.0– 4.9	5
5.0 – 5.9	8
6.0 – 6.9	12
7.0 – 7.9	15
8.0 – 8.9	6
9.0 – 9.9	3

Compute the following:

- a. Range
- b. Mean absolute deviation
- c. Variance
- d. Standard deviation

10. The record of dental clinic inside the mall shows patients that availed of tooth restoration during weekdays:

Day	Monday	Tuesday	Wednesday	Thursday	Friday
No. of Patients	6	10	12	15	16

Compute the variance in the number of patients of dental clinic using two methods.