BUSINESS STATISTICS

CHAPTER 2

CLASSIFICATION OF DATA

CHAPTER 2

TERMS

- Raw data: it is the collected data which have not been organized numerically.
- Arrays: An array is the arrangement of raw numerical data in ascending or descending order of magnitude.
- ***Variable**:* A variable is any quantity or attribute whose value varies from one unit to another. A variable may be classified as;
- i. *Continuous Variable* – A variable which takes all values within a given range
- ii. *Discrete Variable* – A variable whose values are countable.
- ***Frequency***: It is the number of occurrence of a given value or group.
- ***Frequency Distribution***: A tabular arrangement of data by classes together with the corresponding class frequencies is called a frequency distribution (table). It may be:

i. *Ungrouped,*

ii. *Grouped or*

iii. *Categorical.*

UNGROUPED FREQUENCY DISTRIBUTION

- This is the set of possible values of the variable (discrete) together with the associated frequencies.
- When solving ungrouped frequency distribution questions the following steps are used.
  1. List all values of the variable in ascending order of magnitude,
  2. Form a tally column
  3. Sum the tallies to obtain the frequencies of each variable, the frequencies sum up to the total number of observations.

UNGROUPED FREQUENCY DISTRIBUTION

- Example.

The following is a record of number of absentees per day from a factory over 21 days.

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

SOLUTION

NO. OF ABSENTEES	TALLY	NO. OF DAYS. (FREQUENCY)
0	////	4
1	~~////~~ //	7
2	////	4
3	//	2
4	////	4
TOTAL		21

GROUPED FREQUENCY DISTRIBUTION

- This is a set of class intervals for the variable (continuous) together with the associated frequencies.
- **Class interval**: It is a subdivision of the total range of values a variable (i.e. continuous) may take. The groups into which the values are put are called „classes‟ e.g. 41 – 49 or 50 – 59
- **Class limits**: These are end points of the class interval. The end numbers 41 and 50 are called the lower class limit and the numbers 49 and 59 are called the upper class limits.
- **Class Frequency**: It is the number of variables which fall in a given interval.

GROUPED FREQUENCY DISTRIBUTION

- **Class Boundary:** These are the lower and the upper values of a class that mark common points between classes.

To find the class boundaries, we subtract the upper class limit of the first class from the lower class limit of the second class (i.e.50 – 49 = 1)

Then we divide the result by 2, (i.e. 1/2 = 0.5).

We then subtract 0.5 from all the lower class limits and add 0.5 to all the upper class limits to obtain the lower class boundaries and the upper class boundaries respectively.

Hence, simply, it is the dividing line between any two successive classes. Thus the class boundary for the class 41 – 49 is given as 40.5 – 49.5.

GROUPED FREQUENCY DISTRIBUTION

- **Class Size:** This is the difference between the upper and lower class boundaries of a class interval.

i.e. Class Size = Upper class boundary - lower class boundary

For example the class size or width for the class 41 – 49 is given as 49.5 – 40.5 = 9

- **Class Mark (Class Midpoint):** It is obtained by adding the lower and upper class limits and dividing the result by 2. It can also be obtained by adding the lower and upper class boundaries of a class and dividing the result by 2. It is the midpoint of a class interval.

Ie. **Class mark** = (Lower class limit+ Upper class limit) / 2

GROUPED FREQUENCY DISTRIBUTION STEPS

- Find the highest and lowest values.
- Find the range.(highest value – lowest value)
- Select the number of classes desired.

  (sturges rule = 1+3.322(log_{10}(n))

  (where n stands for the number of data values)
- Find the width by dividing the range by the number of classes and rounding up.
- Select a starting point (usually the lowest value); add the width to get the lower limits.
- Find the upper class limits.
- Find the boundaries.
- Tally the data, find the frequencies, and find the cumulative frequency.

EXAMPLE 2.2

The following are the marks obtained by 20 students in an examination:

60	45	72	55	42	65	54	68	74	50
70	58	48	35	64	51	52	60	58	75

Draw a grouped frequency distribution table.

Solution

CLASS LIMITS	TALLY	FREQUENCY
35 - 41	/	1
42 – 48	///	3
49 – 55	~~////~~	5
56 – 62	////	4
63 – 69	///	3
70 - 76	////	4
Total		**20**

EXAMPLE 2.3

Construct a frequency distribution of the data below

1	2	6	7	12
2	6	9	5	13
18	7	3	15	15
17	1	14	5	4
4	16	4	5	8
5	18	5	2	6
9	11	12	1	9
10	11	4	10	2
9	18	8	8	4
7	3	2	6	14

SOLUTION

CLASS LIMIT	TALLY	CLASS BOUNDARY	FREQUENCY
1 – 3	~~////~~ ~~////~~	0.5 – 3.5	10
4 – 6	~~////~~ ~~////~~ ////	3.5 – 6.5	14
7 – 9	~~////~~ ~~////~~	6.5 – 9.5	10
10 – 12	~~////~~ /	9.5 – 12.5	6
13 – 15	~~////~~	12.5 – 15.5	5
16- 18	~~////~~	15.5 – 18.5	5
TOTAL			**50**

CATEGORICAL FREQUENCY DISTRIBUTION

The categorical frequency distribution is used for data that can be placed in specific categories, such as nominal or ordinal level data. For example data such as political affiliation, religious affiliation or major field of study, etc., could be put in a categorical frequency distribution.

Example 2.4

Twenty-five army inductees were given a blood test to determine their blood type. The data set is as follows:

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

Construct a frequency distribution for the data.

Solution

CLASS	TALLY	FREQUENCY	PERCENTAGE (f/n)
A	~~////~~	5	20
B	~~////~~ //	7	28
O	~~////~~ ////	9	36
AB	////	4	16
TOTAL		**25**	**100**

Example 2.5

These data represent the record of high temperatures for each of the 50 states. Construct a frequency distribution for the data using 7 classes.

112	100	127	120	134
118	105	110	109	112
110	118	117	116	118
122	114	114	105	109
107	112	114	115	118
117	118	122	106	110
116	108	110	121	113
120	119	111	104	111
120	115	120	117	105
116	118	112	114	114

Solution

CLASS LIMITS	**CLASS BOUNDARIES**	**TALLY**	**FREQUENCY**	**LESS THAN CUMULATIVE FREQUENCY**
100 – 104	99.5 – 104.5	//	2	2
105 – 109	104.5 – 109.5	~~////~~ //	8	10
110 – 114	109.5 – 114.5	~~////~~ ~~////~~ ~~////~~ ///	18	28
115 – 119	114.5 – 119.5	~~////~~ ~~////~~ ///	13	41
120 – 124	119.5 – 124.5	~~////~~ //	7	48
125 – 129	124.5 – 129.5	/	1	49
130 - 134	129.5 – 134.5	/	1	50

GRAPHICAL REPRESENTATION

The information provided by a frequency distribution in tabular form may be presented graphically by:

- Histogram
- Frequency Polygon
- Cumulative Frequency curve

HISTOGRAM

It contains rectangles having

- Bases on a horizontal axis, centres at the class marks and lengths equal to the class interval sizes
- Areas proportional to class frequencies

EXAMPLE 2.6

- Draw a histogram for the following data:

CLASS INTERVAL	10 - 19	20 - 29	30 - 39	40 - 49	50 – 59
FREQUENCY	2	4	6	2	1

CLASS INTERVAL	CLASS BOUNDARY	FREQUENCY
10 – 19	9.5 – 19.5	2
20 – 29	19.5 – 29.5	4
30 – 39	29.5 – 39.5	6
40 – 49	39.5 – 49.5	2
50 – 59	49.5 – 59.5	1

HISTOGRAM

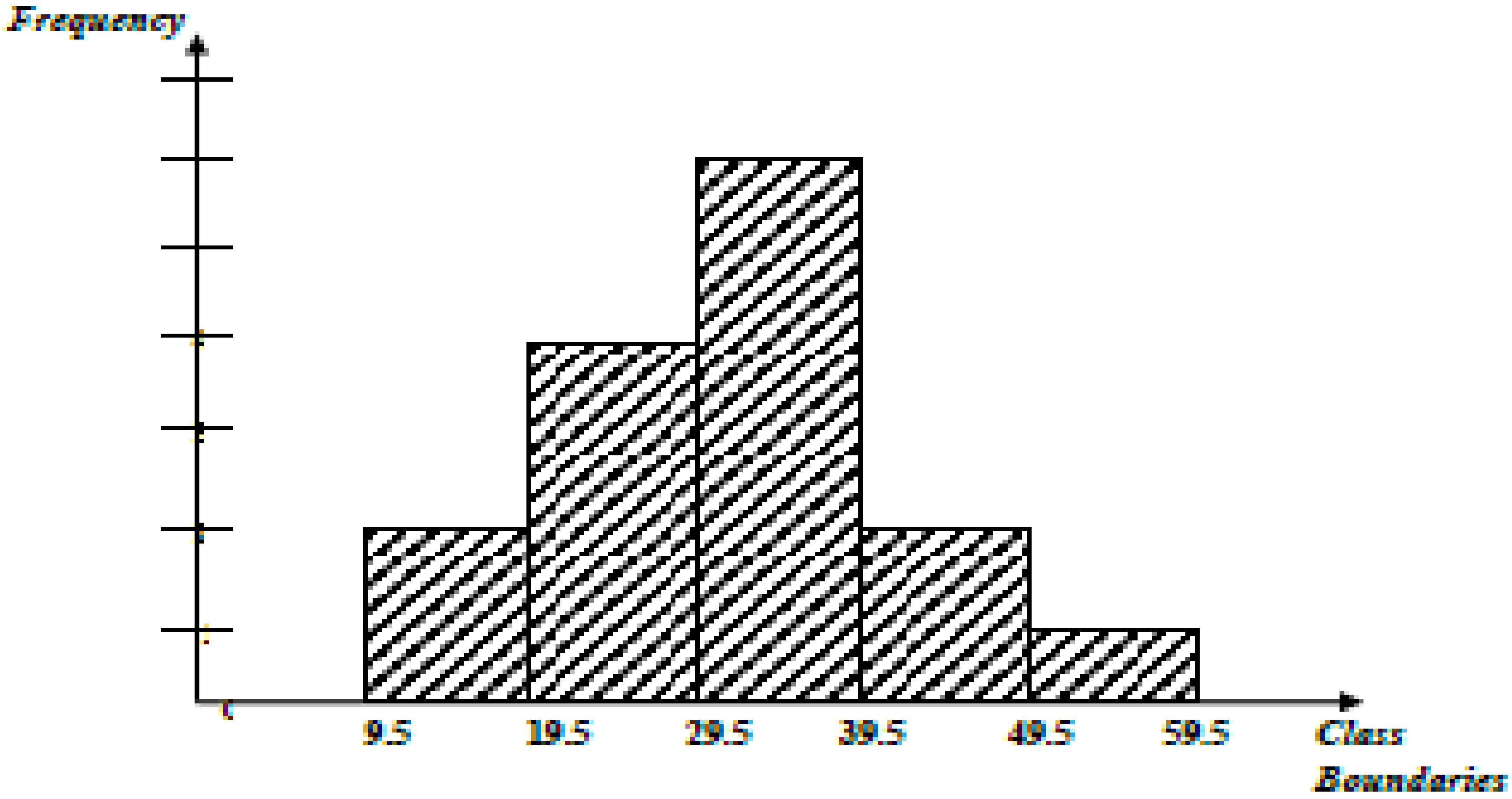


Figure 2.1

FREQUENCY POLYGON

This is a line graph of class frequency against the midpoints. It can be obtained by connected midpoints of the tops of the rectangles in the histogram.

To complete the polygon, the midpoints at each end are joined to the immediate lower or higher midpoints at zero frequency.

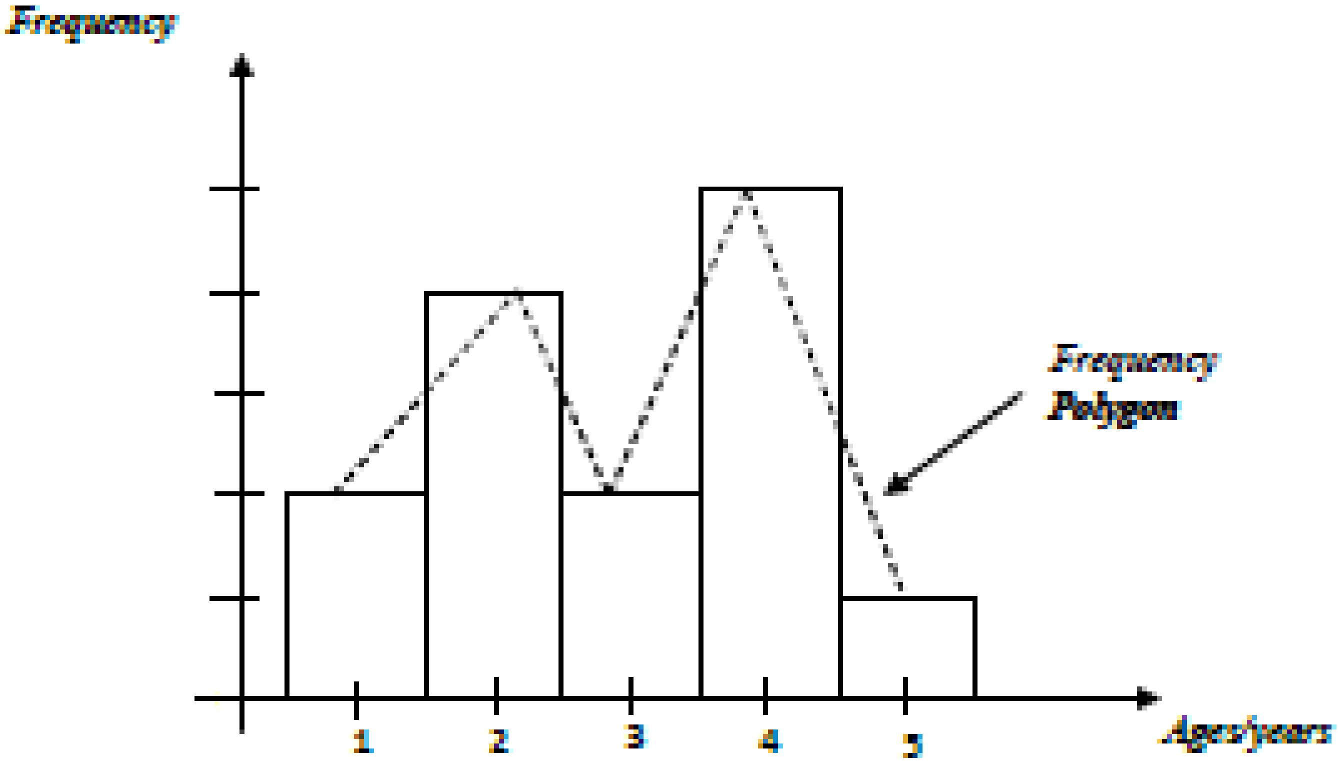
- Draw a histogram for the following data:

AGE (midpoint)	1	2	3	4	5
FREQUENCY	2	4	2	5	1

SOLUTION

CLASS MID-POINT	CLASS BOUNDARY	FREQUENCY
1	0.5 – 1.5	2
2	1.5 – 2.5	4
3	2.5 – 3.5	2
4	3.5 – 4.5	5
5	4.5 – 5.5	1

Graph



CUMULATIVE FREQUENCY

The cumulative frequency corresponding to a class is the sum of frequencies of that class and of all classes preceding that class.

a. A table showing cumulative frequency is called the cumulative frequency table;

b. A graph of cumulative frequency against the corresponding variable is called the cumulative frequency curve or Ogive.

: Unless otherwise stated a question will normally refer to the less than Ogive.

Less than ogives are drawn using the higher class boundaries

More than ogives are drawn using the lower class boundaries

EXAMPLE 2.9

- Draw the ogives for the following distribution of marks obtained by 59 students:

MARKS	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 – 70
FREQ.	4	8	11	15	12	6	3

SOLUTION

MARKS	FREQUENCY	LESS THAN CUMULATIVE FREQUENCY	MORE THAN CUMULATIVE FREQUENCY
0 – 10	4	4	59
10 – 20	8	12	55
20 – 30	11	23	47
30 - 40	15	38	36
40 – 50	12	50	21
50 – 60	6	56	9
60 - 70	3	59	3

EXPLORATORY DATA ANALYSIS – THE STEM AND LEAF DISPLAY

The techniques of exploratory data analysis consist of simple arithmetic and easy-to-draw graphs that can be used to summarize data quickly. One technique – referred to as a stem-and-leaf display – can be used to show both the rank order and shape of a data set simultaneously.

To develop a stem-and-leaf display,

1. we first arrange the leading digits of each data value to the left of a vertical line.
2. To the right of the vertical line,
3. we record the last digit for each data value as we pass through the observations in the order they were recorded.
4. The last digit for each data value is placed on the line corresponding to its first digit.

EXAMPLE

- TABLE: NUMBER OF QUESTIONS ANSWERED CORRECTLY ON AN APTITUDE TEST

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

SOLUTION

	Leaf unit = 1
6	8 9
7	2 3 3 5 6 8
8	0 1 1 2 3 4 5 6
9	1 2 2 2 4 5 5 6 7 8 8
10	0 0 2 4 6 6 6 7 8
11	2 3 5 5 8 9 9
12	4 6 7 8
13	2 4
14	1

Example

Consider the following data on the number of hamburgers sold by a fast-food restaurant for each of 15 weeks:

1565	1852	1644	1766	1888
1912	2044	1812	1790	1679
2008	1852	1967	1954	1733

SOLUTION

	Leaf unit = 10
15	6
16	4 7
17	3 6 9
18	1 5 5 8
19	1 5 6
20	0 4