Chapter 2

Methods of Data Collection and Presentation

Alemakef Wagnew M.(Bsc. in statistics and MPH)

May 8, 2019

2.1 Methods of data collection:

- **Source of Data:**
  Statistical data may be obtained from two sources, namely, primary and secondary
  1. **Primary data:**

     data measured or collected by the investigator or the user directly from the source. Primary sources are sources that can supply first hand information for immediate user.
  2. **Secondary data:**

     When an investigator uses data, which have already been collected by others, such data are called secondary data. Data gathered or compiled from published and unpublished sources

Chapter 2: Methods of data collection

We distinguish three basically different methods of collecting data. These are

1. Extraction of data from records
2. Self-administered questionnaire
3. Direct investigation-measurement (observation) of the subject and interviewing (face-to-face, telephone) our first step is to decide on which of these three methods to use

Methods of data collection

Extractions of data from Records

A mass of information about the population studied by social surveys is available in

- historical documents,
- statistical reports,
- records of institutions and
- other sources

Methods of data collection

Self-administered questionnaire

Self-administered questionnaire : is a method of data collection in which researchers can give questionnaires with instructions directly to respondents or mail them to respondents who read instructions and questions, then record their answers and give it back or return it by mail again to data collecting agency

Advantage:

- It is the cheapest and can be conducted by a single researcher.
- A researcher can send questioners to a wide geographical area.
- The respondent can complete the questionnaire when it is convenient and check personnel records if necessary.
- Mail questionnaire offer anonymity and avoid interviewer bias.
- They are very effective, and response rates may be high for a target population that is well educated or has a strong interest in the topic

Methods of data collection

Disadvantage

- A low response rate is the biggest problem.
- A researcher cannot control the conditions under which a mail questionnaire is completed.
- Researcher cannot visually observe the respondent's reactions to questions, physical characteristics, or settings.
- Mail questionnaire is not suitable for illiterate community.

Methods of data collection

Direct investigation - measurement

These are:

- **Observation and**
- **Interviewing (face-to-face, telephone)**

  Measurement or Observations
  Information on a topic can be gathered by measurement if it is physically measurable or observable. Common types of data collected by observation and measurement include:
  - Land area measurement
  - Crop output measurement
  - Anthropomorphic measurements
  - Animal weight gain
  - Instrument recording or readings (e.g. rainfall, temperature, etc)
  - Physical measurement or examination of people
  - Counts of human, animal and plant populations
  - Direct observations of work

Methods of data collection

Measurement Or Observation

- Exchange activities (e.g. purchases and sale prices).
- Data collection by measurement can be undertaken is several ways, some of these are:
  - The direct measurement of a physical characteristic using an instrument
  - The observation of people engaged in an activity; and
  - Recording of relevant aspects of their activities
- Interviewing (face-to-face, telephone)
  - Face-to-face interview is a social process that involves the interviewer and respondent.
  - It is the process in which the interviewer meets the respondents, explains the purpose of the study, forwards a set of questions and records the answers.
  - It is widely used in economic and social surveys

Methods of data collection

Advantages of face-to-face interviews

- Have the highest response rate and permit the longest questionnaires.
- Interviewers control the sequence of questions and use some probes.
- Respondent is likely to answer all the questions alone.
- Interviewers also can observe the surroundings and can use nonverbal communication and visual aids.
- Well-trained interviewers can ask all types of questions including complex questions.

Methods of data collection

Disadvantage of face to face

- Cost is high- that is , recruiting, training, travel, supervisor, and personnel costs for interviewers can be high.
- Interviewer bias is also high in this method.
- The appearance, tone of voice, question wording, and so forth of the interview may affect the respondent

Methods of data Presentation

1. **Textual Method:** – a narrative description of the data gathered
2. **Tabular Method or frequency distribution :–** a systematic arrangement of information into columns and rows
3. **Graphical Method :–** an illustrative description of the data

THE FREQUENCY DISTRIBUTION TABLE (FDT)

- A FDT is a statistical table showing the frequency or number of observations contained in each of the defined classes or categories.

  Parts of a frequency distribution table:
  - Heading
  - Body
  - Stubs or classes
  - Caption

- **Frequency distribution:** is a basic techniques that provide rich insights into the data and lay the foundation for more advanced analysis.
- **A frequency distribution table:** lists categories of scores along with their corresponding frequencies.

Frequency distribution

- It is a grouping of all the (numerical) observations into intervals or classes together with a count of the number of observations that fall in each interval or class.

- **A frequency distribution has two main parts:**
  - The values of the variable (if quantitative) or the categories (if qualitative), and
  - The number of observations (frequency) corresponding to the values or categories.

Frequency Distributions

There are two types of Frequency distributions

1. Categorical (or qualitative)
2. Numerical (or quantitative)
   - **Categorical Frequency Distribution:**
     - Data are classified according to non-numerical categories.
     - Categories must be mutually exclusive and exhaustive.
     - Used to present nominal and ordinal data

☐ **Nominal data:**Here the construction is straight forward: count the occurrences in each category and find the totals.
Example: The martial status of 60 adults classified as single, married, divorced and widowed is presented in a FD as below:

Marital status	Single	Married	Divorced	Widowed	Total
Frequency	25	20	8	7	60

Categorical Frequency distribution

- **Ordinal data:**The construction is identical to the nominal case. How ever, the categories should be put in an ordered manner

  Example: Satisfaction of hospital admission in Gondar hospital size of 80 is presented a FD as shown below.

Satisfaction	V.Satisfied	Satisfied	Dissatisfied	V.Dissatisfied	Total
Frequency	15	36	3	7	80

Numerical Frequency Distribution

- data are classified according to numerical size.
- used to summarize interval and ratio data.
- may be discrete or continuous, depending on whether the variable is discrete or continuous Population

1. **Discrete (Ungrouped) Frequency Distribution**
   Count the number of times each possible value is repeated
   Example: In a survey of 30 families, the number of children per family was recorded and obtained the following data:

   4 2 4 3 2 8 3 4 4 2 2 8 5 3 4 5 4 5 4 3 5 2 7 3 3 6 7 3 8 4 These

   individual observations can be arranged in ascending order of magnitude to from an array:
   2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4 4 4 4 5 5 5 5 6 7 7 8 8 8
   The distribution of children in 30 families would be:

Number of children	2	3	4	5	6	7	8	Total
Number of family	5	7	8	4	1	2	3	30

Continuous frequency distribution

The purpose of this kind of a table is to select a set of intervals on the number line, then count the number of values that fall into each interval

definitions of some basic terms

- **Range (R)** – the difference between the highest score and the lowest score.
- **Class Interval (k)**– a grouping or category defined by a lower limit and an upper limit.
- **Class Boundaries (CB)** – these are also known as the exact limits, and can be obtained by subtracting 0.5 from the lower limit of an interval and adding 0.5 to the upper limit interval.

Continuous frequency distribution

continuous or grouped frequency distribution

- **Class Mark (x)** – is the middle value or the midpoint of a class interval. It is obtained by getting the average of the lower class limit and the upper class limit
- **Class Size or width (W)** – is the difference between the upper class boundary and the lower class boundary of a class interval
- **Relative Frequency (RF)** – these are the percentage distribution in every class interval.
- **Class Frequency** – it refers to the number of observations belonging to a class interval, or the number of items within a category

Steps in Constructing a (FDT)

Example: The blood glucose level for 50 patients is shown below. Construct a frequency distribution for the following data 51 65 68 87 76 56 69 75 89 80 61 66 73 86 79 70 71 54 87 78 68 74 66 88 77 67 73 64 90 77 72 52 67 86 79 74 59 70 89 85 55 63 74 82 84 57 68 72 81 83

- **Step 1:**Find highest and lowest value and compute range R, using the formula:

$$R = HighestScore–LowestScore \tag{1}$$

- **Step 2:**Compute for the number of class intervals, K, by using the formula **Decide k with the help of Sturge's rule**

$$k = 1 + 3.3\ logn \tag{2}$$

Note: The ideal number of class intervals should be 5 to 15. Less than 8 intervals are recommended for a data with less than 50 observations/values. For a data with 50 to 100 observations/values, the suggested number should be greater than 8

Steps in Constructing a (FDT)

Example: The blood glucose level for 50 patients is shown below. Construct a frequency distribution for the following data

- **Step 3:**Find the class width; w=R/k and closest ones (rounding up)

$$w = R/k \tag{3}$$

- **Step 4**:Select the starting observation as lowest class limit (this is usually the lowest observation). Add the width to that observation to get the lower limit of the next class. Keep adding until there are k classes
- **Step 5:** Find the upper class limit
- **Step 7:** Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to the UCL as shown.
- **step 8:**Tally the data
- **step 9:** Write the numeric values for the frequency column
- **Step 10:** Find cumulative frequency.
- **Step 11:** Find relative frequency and /or relative cumulative frequency.

Solution of Example:

i. **Step 1:** Highest value 90 and lowest 51 then

$$range = 90 - 51, range = 39 \tag{4}$$

ii. **Step 2:** the class limit using Sturge's rule

$$k = 1 + 3.3 \log 50, k = 8 \tag{5}$$

iii. **Step 3:**Find the class width

$$w = r/k, w = 39/8, w = 4.875 \approx 5 \tag{6}$$

iv. **Step 4:**Select the starting observation as lowest class limit and Add the width to that observation to get the lower limit of the next class. Keep adding until there are 5 classes 51, 56, 61, 66, 71, 76, 81, 86, 91

v. **Step 5:**Find the upper class limit; e.g.

$$thefirstupperclass = 51 - (U = 56 - 1) = 55 \tag{7}$$

55, 60, 65, 70, 75, 80, 85 and 90 will be the upper class limit

Example solution

So combining step 4 and step 5, one can construct the following classes: So combining step 5 and step 6, one can construct the following classes.

class limit
51-55
56-60
61-65
66-70
71-75
76-80
81-85
86-90

- **Step 6:**: Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to the UCL as shown **Example:**

$$LcB = 51 - 0.5 = 50.05, andUcB = 55 + 0.5 = 55.05 \quad (8)$$

Example solution

Then continue adding W on both boundaries to obtain the rest boundaries. By doing so one can obtain the following classes

class boundary
50.5 - 55.5
55.5 - 60.5
60.5 - 65.5
65.5 - 70.5
70.5 - 75.5
75.5 - 80.5
80.5 - 85.5
85.5 - 90 .5

Step 7: Tally the data
step 8: Write the numeric values for the tallies in the frequency column
Step 9: Find cumulative frequency relative frequency and /or relative cumulative frequency

The complete frequency distribution table

C.limit	C.boun	C. M	Freq.	<CF	> *CF*	RF	<RCF	>RCF
51-55	50.5-55.5	53	4	4	50	0.08	0.08	1
56-60	55.5-60.5	58	3	7	43	0.06	0.14	0.86
61-65	60.5-65.5	63	4	11	39	0.08	0.22	0.78
66-70	65.5-70.5	68	10	21	29	0.2	0.42	0.58
71-75	70.5-75.5	73	9	30	20	0.18	0.60	0.40
76-80	75.5-80.5	78	7	37	13	0.14	0.74	0.26
81-85	80.5-85.5	83	5	42	8	0.1	0.84	0.16
86-90	85.5-90.5	88	8	50	0	0.16	1	0

Continuous frequency distribution

Example 2:

Construct a continuous FD for the following raw data of ages of patients admitted at felege hiwot hospital in a given week.

57, 53, 65, 55, 50, 45, 64, 52, 16, 46, 42, 63, 33, 64, 53, 25, 54, 35, 48, 55, 70, 47, 39, 58, 52, 36, 65, 75, 26, 20, 55, 60, 83, 61, 45, 63, 49, 42, 35, 18, 51, 45, 42, 65, 39, 59, 45, 41, 30, 40

Class work

Diagrammatic and Graphical Methods of Data Presentation

A F.D can be presented graphically or diagrammatically

Advantage:

I. To understand the information easily.

II. To make the data attractive.

III. To make comparisons of items easy.

IV. To draw attention of the observer

- The purpose of graphs and diagrams is not to provide exact and detailed information, but simple comparisons. Any further information shall rather be obtained from the original data.

2.2.2 Diagrammatic Presentation of Data

Diagrams are appropriate for presenting discrete as well as qualitative data. The three most commonly used diagrammatic presentation of data are:

- Pie charts
- Bar charts
- Pictograms

Pie Chart

- **Pie chart:** can used to compare the relation between the whole and its components
- **Pie chart:** is important for depicting discrete variables with relatively few categories.
- **Pie chart:** is a circular diagram and the area of the sector of a circle is used in pie chart.

Steps in constructing a pie-chart

Steps in constructing a pie-chart

1. Construct a frequency table
2. Change the frequency into percentage (P) or fraction (F)
3. Change the percentages into degrees, where:

$$\text{Angle of Sector} = \frac{componentpart}{total} * 360^{\circ} \tag{9}$$

4. Draw a circle and divide it accordingly

pie Chart

Example Pie chart:

The following table gives the details of monthly budget of a family. Represent these figures by a suitable diagram.

item of expenditure	family budget
Food	600
Clothing	100
house rent	400
fuel and lighting	100
miscellaneous	300
Total	1500

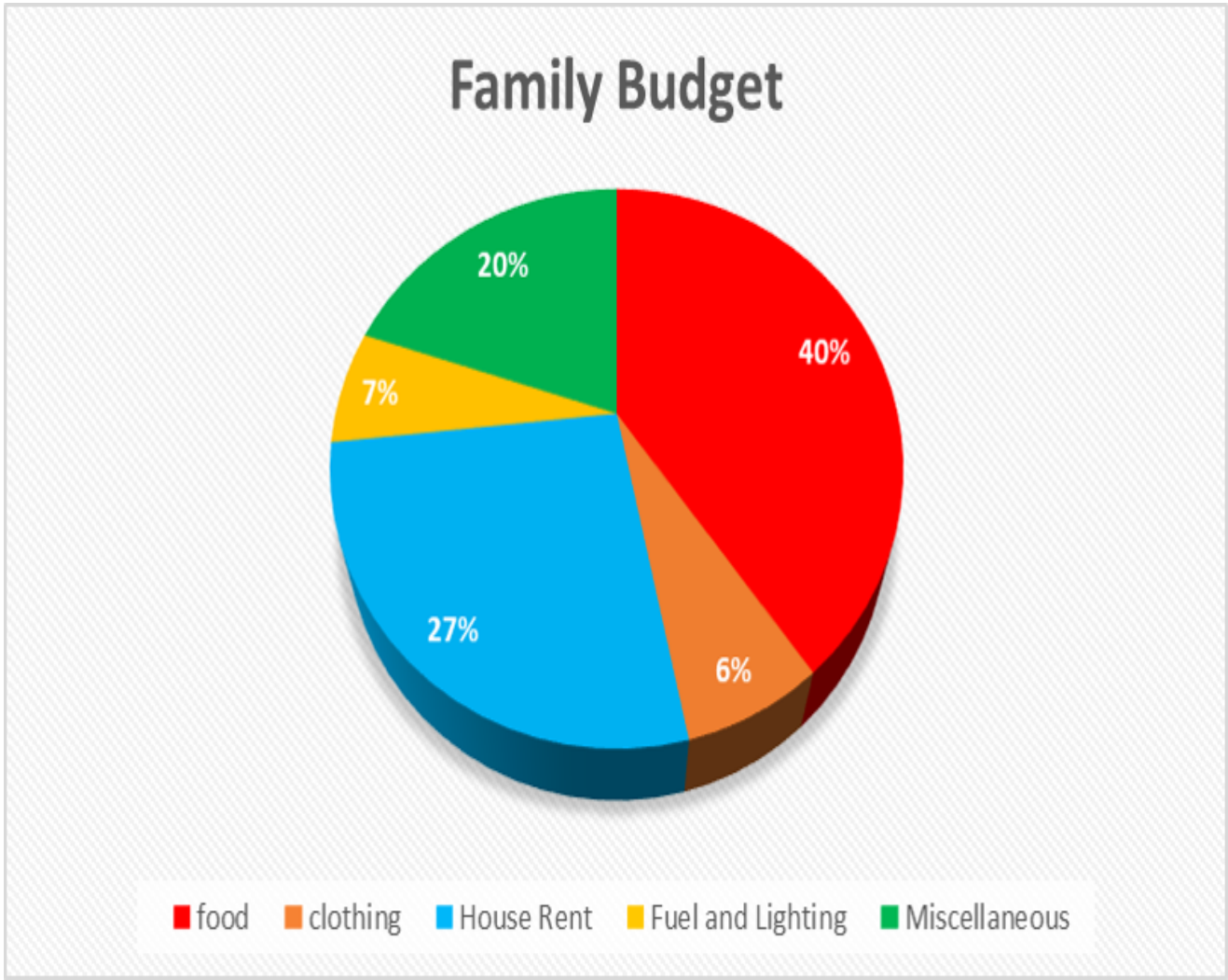
Pie chart

Example:

continued

item of expenditure	family budget	Angel of sector	Percent
Food	600	144°	40
Clothing	100	24°	6.67
house rent	100	96°	26.67
fuel and lighting	100	24°	6.67
miscellaneous	300	72°	20
Total	1500	360°	100

Pie Chart Example



Bar chart

The bar chart (simple, multiple and stacked bar graph) :

used to represent and compare the frequency distribution of discrete variables and attributes or categorical series.
The vertical or horizontal bins to represent the frequencies of a distribution. While we draw bar chart, we have to consider the following points. These are (see the following slide)

Tips for constructing bar chart:

I. Whenever possible it is better to construct a bar diagram on a graph paper
II. All bars drawn in any single study should be of the same width
III. The different bars should be separated by equal distances
IV. All the bars should rest on the same line called the base
V. Whenever possible, it is advisable to draw bars in order of magnitude

Simple Bar Chart

simple bar Chart: is used to represents data involving only one variable classified on spatial, quantitative or temporal basis.

Example: Draw simple bar diagram to represent the profits of a bank for 5 years.

year	1989	1990	1991	1992	1993
profit(milions$)	10	12	18	25	42

Simple bar Chart

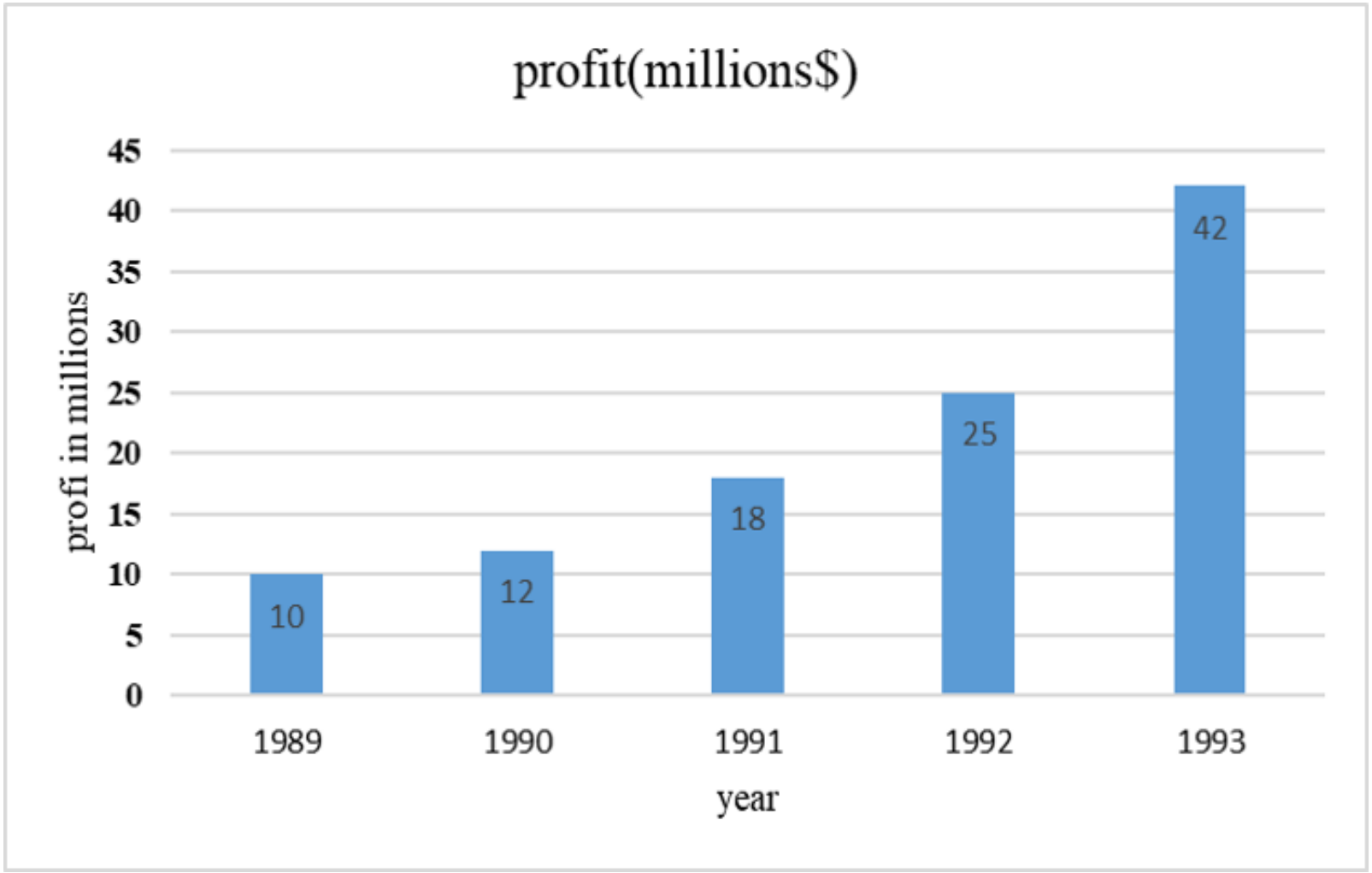


Figure: Simple bar chart

Multiple bar Chart

- Multiple bar chart are used two or more sets of inter-related data are represented (multiple bar diagram facilities comparison between more than one phenomenon).

Example : Draw a multiple bar chart to represent the import and export of Canada (values in $) for the years 1991 to 1995.

year	imports	Exports
1991	7930	4260
1992	8850	5225
1993	9780	6150
1994	11720	7340
1995	12150	8145

Multiple bar Chart Example

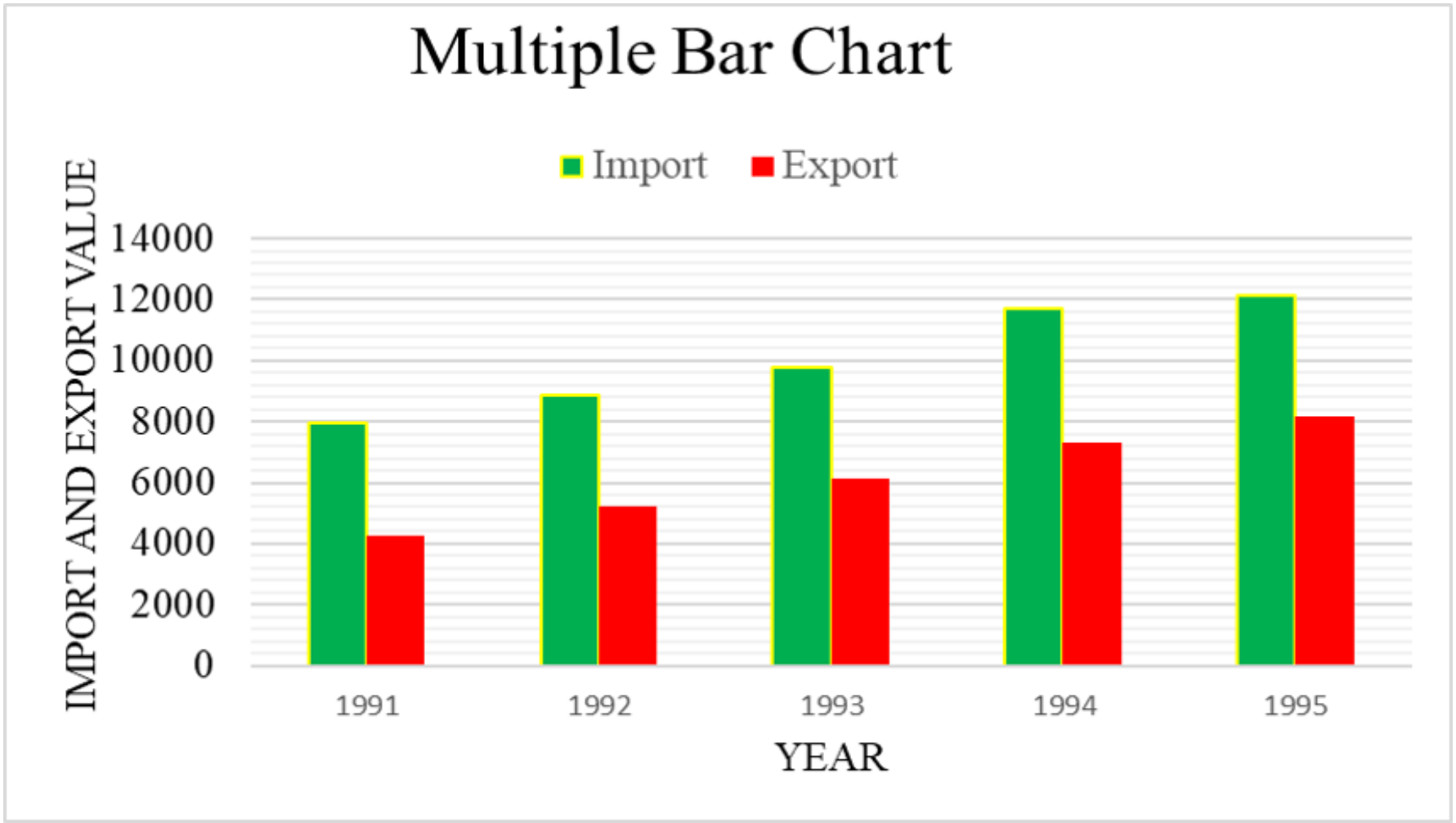


Figure: Multiple bar chart

Component Bar Chart

Component bar Chart: is used to represent data in which the total magnitude is divided into different or components

Example : The table below shows the quantity in hundred kgs of Wheat, Barley and Oats produced on a certain form during the years 1991 to 1994.

Draw stratified bar chart:

year	wheat	barely	oats
1991	34	18	27
1992	43	14	24
1993	43	16	27
1994	45	13	34

Component bar Chart

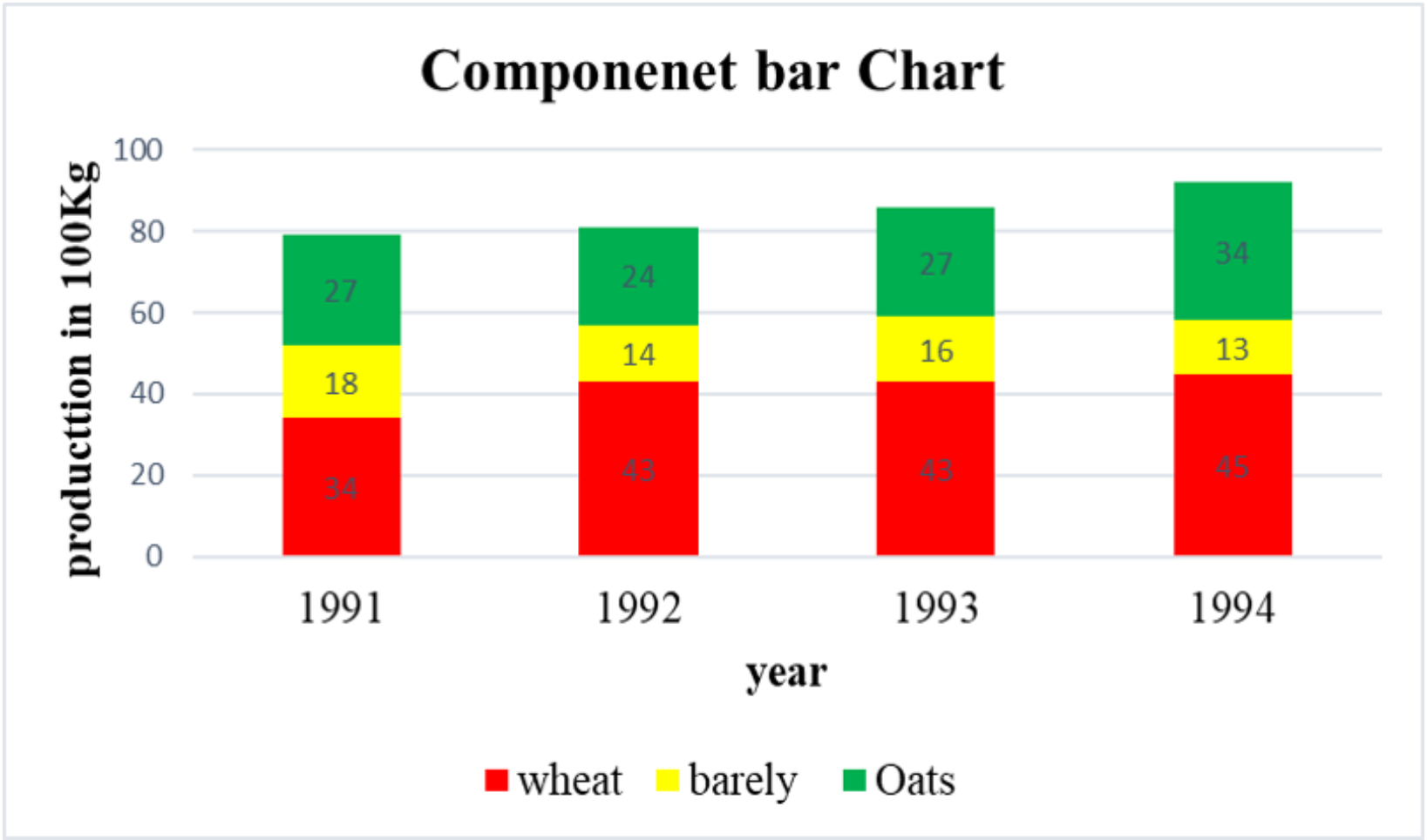


Figure: Component bar chart

Graphical Presentation of data

- The histogram,
- frequency polygon and
- cumulative frequency graph (ogive) are most commonly applied graphical representation for **continuous data**.
  Procedures for constructing statistical graphs
  - Draw and label the X and Y axes.
  - Choose a suitable scale for the frequencies or cumulative frequencies and label it on the Y axes.
  - Represent the class boundaries for the histogram or ogive and the mid points for the frequency polygon on the X axes.
  - Plot the points.
  - Draw the bars or lines to connect the points.

Histogram

I. A graph which places the class boundaries on the horizontal axis and the frequencies on a vertical axis

II. Class marks and class limits are some times used as quantity on the X axes.

III. **Example:** Construct a histogram to by using the following data

The blood glucose level for 50 patients is shown below. Construct a frequency distribution for the following data.

44	50	79	63	66	54	56	70	56	63
60	87	60	70	59	60	62	88	71	53
56	65	74	80	51	83	69	77	69	50
58	42	43	85	43	75	55	60	58	49
72	67	55	77	48	45	61	47	44	61

Example

Class limits	Class boundary	Class Mark	Freq.	<CF	>CF	RF	<RCF	>RCF
42-48	41.5 – 48.5	45	8	8	50	0.16	0.16	1
49-55	48.5 – 55.5	52	8	16	42	0.16	0.32	0.84
56-62	55.5 – 62.5	59	13	29	34	0.26	0.58	0.68
63-69	62.5 – 69.5	66	7	36	21	0.14	0.72	0.42
70-76	69.5 – 76.5	73	6	42	14	0.12	0.84	0.28
77-83	76.5 – 83.5	80	5	47	8	0.10	0.94	0.16
84-90	81.5 – 90.5	87	3	50	3	0.06	1	0.06
Total			50			1		

Figure: Grouped Frequency Distributions of 50 Patients glucose level

Histogram Graph

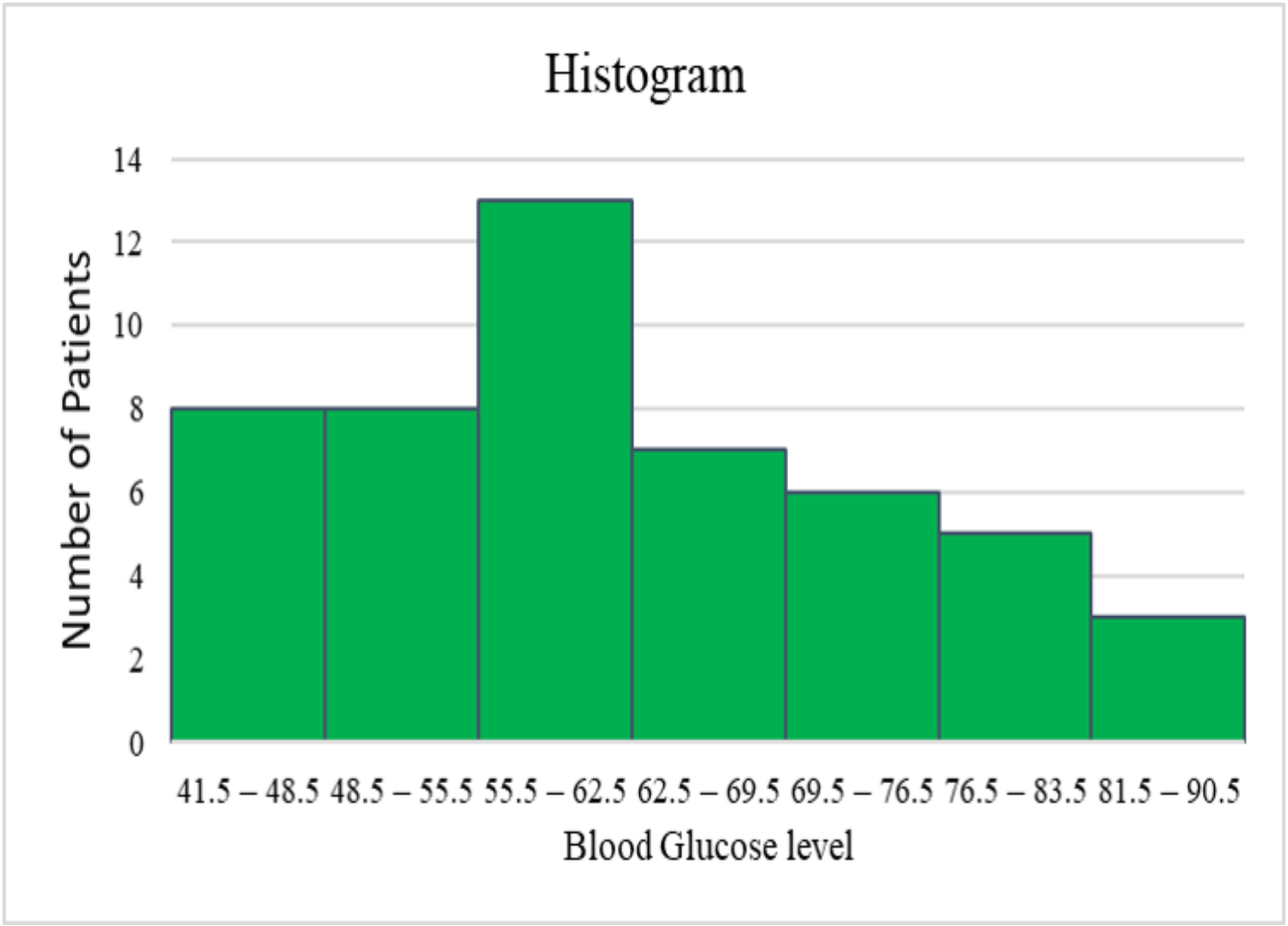
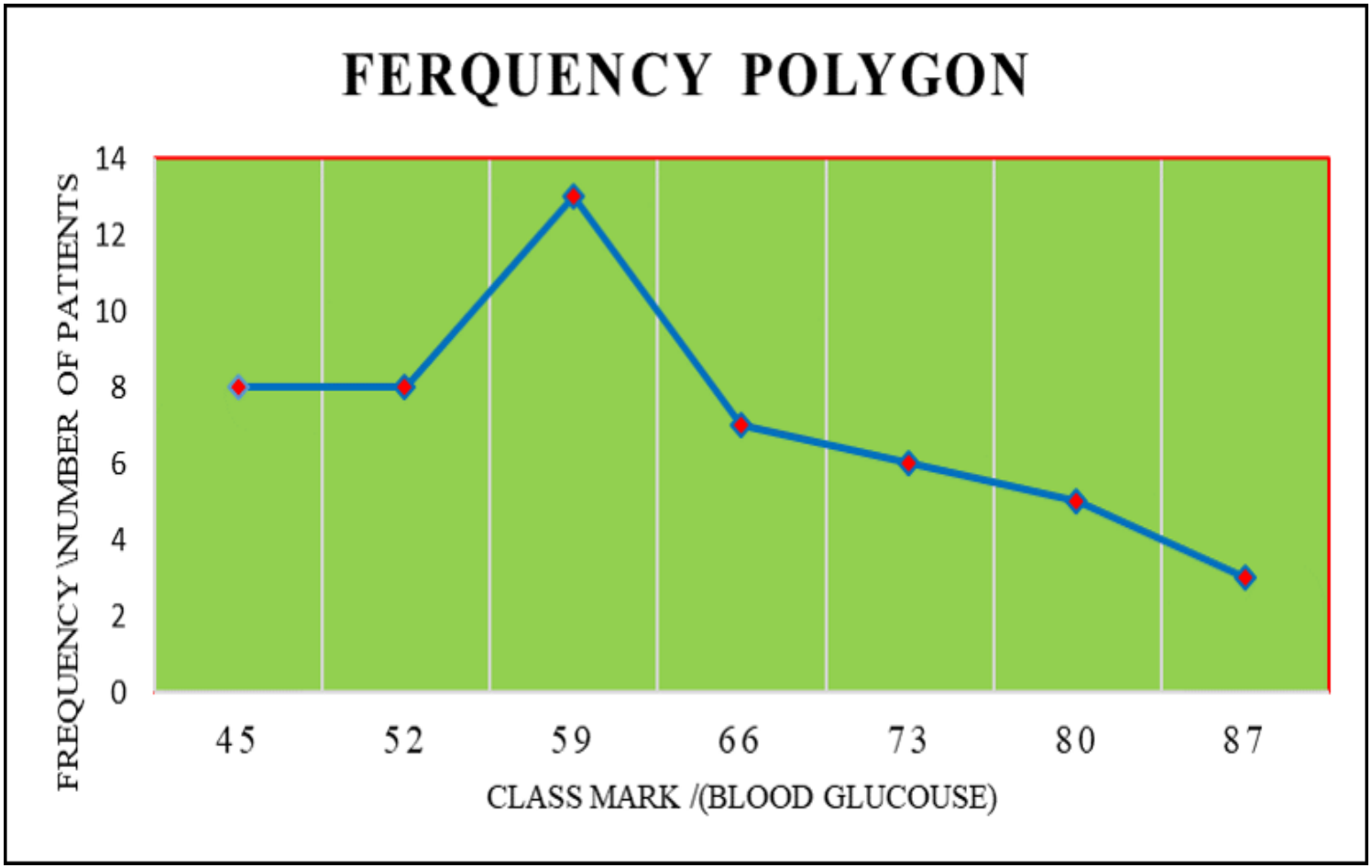


Figure: Histogram Graph of 50 Patients Blood glucose level

Frequency polygon

- Line graph of class marks against class frequencies.
- To draw a frequency polygon we connect the midpoints of class boundaries of the histogram by a straight line



Ogive (cumulative frequency polygon)

- A graph showing the cumulative frequency (less than or more than type) plotted against upper or lower class boundaries respectively.
- That is class boundaries are plotted along the horizontal axis and the corresponding cumulative frequencies are plotted along the vertical axis.
- The points are joined by a free hand curve.
  Example: Draw an ogive curve(less than type) for the above data.

Ogive (cummulative less than type)

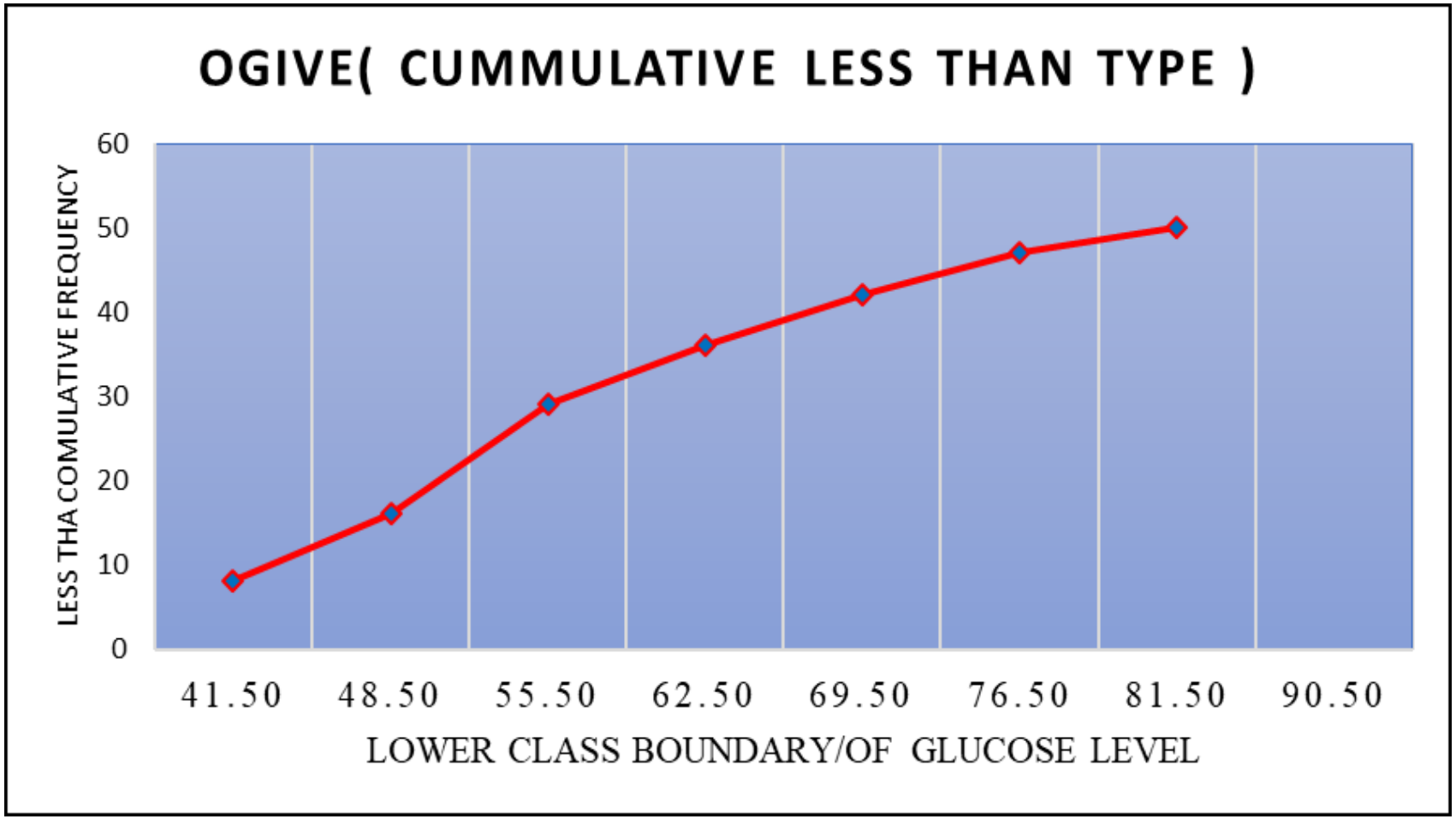


Figure: Ogive (cumulative less than type)

Thank you