

Chapter Two

2. Methods of Data Collection and Presentation

2.1. Methods of Data Collection

Any scientific investigation requires data related to the study. The required data can be obtained from either a **primary** source or a **secondary** source. There are two sources of data:

1. Primary Data

- Data measured or collect by the investigator or the user directly from the source.
- Two activities involved: planning and measuring.

a) Planning:

- Identify source and elements of the data.
- Decide whether to consider sample or census.
- If sampling is preferred, decide on sample size, selection method,... etc
- Decide measurement procedure.
- Set up the necessary organizational structure.

b) Measuring: there are different options.

- Focus Group
- Telephone Interview
- Mail Questionnaires
- Door-to-Door Survey
- Mall Intercept
- New Product Registration
- Personal Interview and
- Experiments are some of the sources for collecting the primary data.

2. Secondary Data: are individuals or agencies, which supply data originally collected for other purposes by them or others.

- Data gathered or compiled from published and unpublished sources or files.
- When our source is secondary data check that:
 - The type and objective of the situations.
 - The purpose for which the data are collected and compatible with the present problem.
 - The nature and classification of data is appropriate to our problem.
 - There are no biases and misreporting in the published data.

Note: Data which are primary for one may be secondary for the other.

2.2. Methods of Data Presentation

After having the collected and edited the data, the next important step is to organize it. That is to present it in a readily comprehensible condensed form

that aids in order to draw inferences from it. It is also necessary that the like be separated from the unlike ones.

The presentation of data is broadly classified in to the following two categories:

- Tabular presentation
- Diagrammatic and Graphic presentation.

The process of arranging data in to classes or categories according to similarities or differences is called *classification*.

Classification is a preliminary and it prepares the ground for proper presentation of data.

1. Tabular Presentation (Frequency distribution)

Definitions:

- **Raw data:** is a data which is collected in original form (survey), whether it may be counts or measurements.
- **Frequency:** is the number of values in a specific class of the distribution.
- **Frequency distribution:** is the organization of raw data in table form using classes and frequencies.

Example: A frequency distribution presenting the number of males and females in a class

Sex	Frequency
Male	57
Female	39

There are three basic types of frequency distributions

- Categorical frequency distribution
- Ungrouped frequency distribution
- Grouped frequency distribution

There are specific procedures for constructing each type.

NB: The main purpose of grouping is now summarization and condensation of the masses of data.

1) Categorical(Qualitative) frequency Distribution:

Used for data that can be placed in specific categories such as nominal or ordinal data.

E.g. Marital status

Example: A social worker collected the following data on marital status for 25 persons. (M=married, S=single, W=widowed, D=divorced)

M	S	D	W	D
S	S	M	M	M
W	D	S	M	M
W	D	D	S	S
S	W	W	D	D

Solution:

Since the data are categorical, discrete classes can be used. There are four types of marital status M, S, D, and W. These types will be used as class for the distribution. We follow procedure to construct the frequency distribution.

Step 1: Make a table as shown.

Class	Tally	Frequency	Percentage
(1)	(2)	(3)	(4)
M			
S			
D			
W			

Step 2: Tally the data and place the result in column (2).

Step 3: Count the tally and place the result in column (3).

Step 4: Find the percentages of values in each class by using;

$$\% = \frac{f}{n} \times 100$$

Where f = frequency of the class, n = total number of value.

Percentages are not normally a part of frequency distribution but they can be added since they are used in certain types of diagrammatic such as pie charts.

Step 5: Find the total for column (3) and (4).

Combing the entire steps one can construct the following frequency distribution.

Class (1)	Tally (2)	Frequency (3)	Percentage (4)	
M	/	6	24	1)
S	//	7	28	2)
D	//	7	28	3)
W		5	20	

2) Ungrouped frequency Distribution:

- Is a table of all the potential raw score values that could possible occur in the data along with the number of times each actually occurred.
- Is often constructed for small set or data on discrete variable.

Constructing ungrouped frequency distribution:

- First find the smallest and largest raw score in the collected data.
- Arrange the data in order of magnitude and count the frequency.
- To facilitate counting one may include a column of tallies.

Example:

The following data represent the mark of 20 students.

80	76	90	85	80
70	60	62	70	55
65	60	63	74	55
76	70	70	80	85

Construct a frequency distribution, which is ungrouped.

Solution:

Step 1: Find the range, Range=Max-Min=90-60=30.

Step 2: Make a table as shown

Step 3: Tally the data.

Step 4: Compute the frequency.

Mark	Tally	Frequency
60	//	2
62	/	1
63	/	1
65	/	1
70	////	4
74	/	1
75	//	2
76	/	1
80	///	3
85	///	3
90	/	1

Each individual value is presented separately, that is why it is named ungrouped frequency distribution.

3) Grouped frequency Distribution:

- ✓ When the range of the data is large, the data must be grouped in to classes that are more than one unit in width.

Definitions:

- **Grouped Frequency Distribution:** a frequency distribution when several numbers are grouped in one class.
- **Class limits (CL):** Separates one class in a grouped frequency distribution from another. The limits could actually appear in the data and have gaps between the upper limits of one class and lower limit of the next.
- **Units of measurement (U):** the distance between two possible consecutive measures. It is usually taken as 1, 0.1, 0.01, 0.001, -----.
- **Class boundaries:** Separates one class in a grouped frequency distribution from another. The boundaries have one more decimal places than the row data and therefore do not appear in the data. There is no gap between the upper boundary of one class and lower boundary of the next class. The lower class boundary is found by subtracting $U/2$ from the corresponding lower class limit and the upper class boundary is found by adding $U/2$ to the corresponding upper class limit. That is,
 $LCB = LCL - \frac{1}{2}U$ and $UCB = UCL + \frac{1}{2}U$
- **Class width (W):** the difference between the upper and lower class boundaries of any class. It is also the difference between the lower limits

of any two consecutive classes or the difference between any two consecutive class marks.

Class mark (Mid points): it is the average of the lower and upper class limits or the average of upper and lower class boundary. i.e.

$$M = \frac{UCB_i + LCB_i}{2}$$

- **Cumulative frequency:** is the number of observations less than/more than or equal to a specific value.
- **Cumulative frequency above:** it is the total frequency of all values greater than or equal to the lower class boundary of a given class.
- **Cumulative frequency below:** it is the total frequency of all values less than or equal to the upper class boundary of a given class.
- **Cumulative Frequency Distribution (CFD):** it is the tabular arrangement of class interval together with their corresponding cumulative frequencies. It can be more than or less than type, depending on the type of cumulative frequency used.
- **Relative frequency (rf):** it is the frequency divided by the total frequency.

$$\text{Relative frequency of a class} = \frac{\text{Frequency of that class}}{\text{Total frequency}}$$

- **Relative cumulative frequency (rcf):** it is the cumulative frequency divided by the total frequency.

Guidelines for classes

1. There should be between 5 and 20 classes.
2. The classes must be mutually exclusive. This means that no data value can fall into two different classes
3. The classes must be all inclusive or exhaustive. This means that all data values must be included.
4. The classes must be continuous. There are no gaps in a frequency distribution.
5. The classes must be equal in width. The exception here is the first or last class. It is possible to have an "below ..." or "... and above" class. This is often used with ages.

Steps for constructing Grouped frequency Distribution

1. First arrange the data in ascending order and determine the unit of measurement, U

2. Find the largest and smallest values , then Compute the Range(R) = Maximum - Minimum

3. Select the number of classes interval (K) desired, usually between 5 and 20 or use Sturges formula:

$$k=1+3.32\log n$$

where k is number of classes desired and n is total number of observation.

🚩 **NB:** k must be rounded up/down to the nearest whole number.

4. Find the class width: It is the gap between two consecutive class intervals. Dividing the range by the number of classes and rounding

$$w = \frac{R}{k}$$

up.

✓ When the data is given as

- Whole number "**w**" always rounded up to the next whole number.

$$\text{e.g. } w=4.13 \approx 5$$

- With one decimal "**w**" always rounded up to the next 1st decimal.

$$\text{e.g. } w=0.325 \approx 0.4.$$

- With two decimals "**w**" always rounded up to the next 2nd decimal.

$$\text{e.g. } w=2.532 \approx 2.54; \wedge w=0.981 \approx 0.99.$$

5. Find the class limit: They are called lower and upper class limits.

Pick a suitable starting point less than or equal to the minimum value. The starting point is called the lower limit of the first class. Continue to add the class width to this lower limit to get the rest of the lower limits.

- **Lower class limit (LCL):** The LCL of the first class interval should be equal to or smaller than the smallest observation in the data.

$$\text{i.e. } lcl_1 \leq \text{the smallest observation}$$

🚩 Continue to add the class width to this lower limit to get the rest of the lower limits. i.e.

$$lcl_{i+1} = lcl_i + w, i = 1, 2, \dots, k-1.$$

- **Upper class limit (UCL):** To find the upper class limit of the **first class**, subtract u from the lower limit of the **second class**. i.e. $ucl_1 = lcl_2 - u$.

✚ Then continue to add the class width to this upper limit to get the rest of the upper class limits. i.e.
 $ucl_{i+1} = ucl_i + w, i = 1, 2, \dots, k-1$.

6. Find the class boundary: are the set of exact limits or true limits. They are called lower and upper class boundaries.

❖ **Lower class boundary (LCB):** The **Lcb** is obtained by subtracting half the unit of measurements from the **lcl** of the class. i.e.

$$lcb_i = lcl_i - \frac{u}{2} \quad \text{Note: } lcb_{i+1} = lcb_i + w$$

❖ **Upper class boundary (UCB):** The **Ucb** is obtained by adding half the unit of measurements from the **ucl** of the class. i.e.

$$ucb_i = ucl_i + \frac{u}{2} \quad \text{Note: } ucb_{i+1} = ucb_i + w$$

7. **Class marks (mid points) (m):** It is the average of **Lcl** and **Ucl** or **Lcb** and **Ucb**.

$$m_i = \frac{lcl_i + ucl_i}{2} \vee m_i = \frac{lcb_i + ucb_i}{2} \quad \text{Note: } m_{i+1} = m_i + w$$

8. Tally the data.
9. Find the frequencies.
10. Find the cumulative frequencies. Depending on what you're trying to accomplish, it may not be necessary to find the cumulative frequencies.
11. If necessary, find the relative frequencies and/or relative cumulative frequencies

Example: Construct a grouped frequency distribution for the following data.

1	2	6	3	1	3	2	2	1	2
1	9		3	4	1	2	7	9	0
1	1	2	3	2	2	2	3	3	2
8	7	2	8	3	1	6	4	9	7

Solutions:

Step 1: Arrange the data in ascending order and $U=20-19=1$

Step 2: Find the range (R) : $R=Max-Min=39-6=33$.

Step 3: Select the number of classes desired using Sturge's formula;

$$k=1+3.322 \times \log n = k=1+3.322 \times \log(20)=5.32 \approx 5(\text{rounding down}).$$

Step 4: Find the class width;

$$w=\frac{R}{k}=w=\frac{33}{5}=6.6 \approx 7(\text{always rounding up}).$$

Step 5: Find the lower and the upper class limits.

- Select the starting point let it be the smallest observation.
 - 6, 13, 20, 27, 34 are the lower class limits.

- Find the upper class limits; e.g. the first upper class limit *i.e.*

$$ucl_1=13-u=13-1=12.$$

- UCL=12, 19, 26, 33, 40 are the upper class limits.

✚ So combining $lcl \wedge ucl$, one can construct the following classes.

Class limits
6 - 12
13 - 19
20 - 26
27 - 33
34 - 40

Step 6: Find the class boundaries;

$$E.g. \text{ For class 1 ; } lcb_1=6-\frac{u}{2}=6-\frac{1}{2}=5.5 \text{ and}$$

$$ucb_1=12+\frac{u}{2}=12+\frac{1}{2}=12.5$$

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

Class boundary
5.5 - 12.5
12.5 - 19.5
19.5 - 26.5
26.5 - 33.5
33.5 - 39.5

Step 7: Find the frequencies.

✓ The complete frequency distribution is given as follows:

Class limit	Class boundary	Class Mark	f	Lcf	Mcf	rf.	%rf	%rc f
6 - 12	5.5 - 12.5	9	2	$\leq 12.5 (\leq 12) = 2$	$\geq 5.5 (\geq 6) = 20$	0.10	10%	10%
13 - 19	12.5 - 19.5	16	4	$\leq 19.5 (\leq 19) = 6$	$\geq 12.5 (\geq 13) = 18$	0.20	20%	30%
20 - 26	19.5 - 26.5	23	6	$\leq 26.5 (\leq 26) = 12$	$\geq 19.5 (\geq 20) = 14$	0.30	30%	60%
27 - 33	26.5 - 33.5	30	5	$\leq 33.5 (\leq 33) = 17$	$\geq 26.5 (\geq 27) = 8$	0.25	25%	85%
34 - 40	33.5 - 39.5	37	3	$\leq 39.5 (\leq 39) = 20$	$\geq 33.5 (\geq 34) = 3$	0.15	15%	100%

Diagrammatical presentation of data.

These are techniques for presenting data in visual displays using geometric and pictures.

Importance:

- They have greater attraction.
- They facilitate comparison.
- They are easily understandable.

- ❖ Usually diagrams are appropriate for presenting discrete data.
- ❖ The three most commonly used diagrammatic presentation for discrete as well as qualitative data are:

- Pie charts
- pictogram
- Bar charts

1. Pie chart

A pie chart is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution. The angle of the sector of a class is obtained by multiplying the ratio of the frequency of the class to the total frequency by 360° .

$$i.e. \text{ sector angle of a class} = \frac{\text{frequency of the class}}{\text{total frequency}} \times 360^\circ$$

Note that: pie-charts are usually used for depicting nominal level data.

Example: Draw the pie chart for the following hospital data. First construct a table providing the central angles.

How to draw a pie-chart

- First find the percentages of each class
- Next calculate the degree measures for each class
- Finally, using a protractor, put each sector /degree measure/ in a circle and give a key for explanation.

Wards	Frequency	Percentage rf	Central angle
Medical A	85	42.5%	153 ⁰
Surgical A	65	32.5%	117 ⁰
Pediatrics	50	25%	90 ⁰
Total	200	100%	360 ⁰

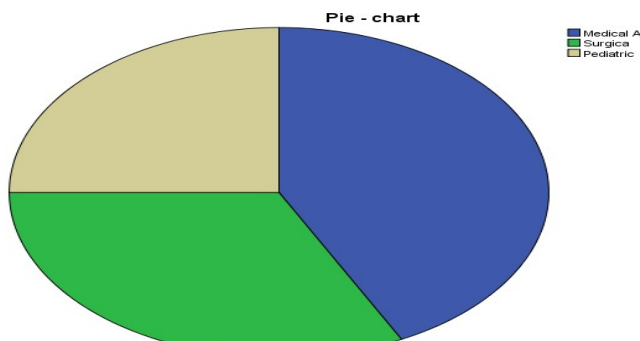


Figure 4: Pie-chart for the above Hospital data.

2. Pictogram

- In pictograms, we represent the data by means of some picture symbols. Here we decide a suitable picture to represent a definite number of units in which the variable is measured.

Example: Draw a pictorial diagram to present the following data (number of students in a certain school for four years.)

Year	1992	1993	1994	1995
No. of students	2000	3000	5000	7000

Let a single picture (□) represents one thousand students.

1995	□□□□□□
1994	□□□□□

Key: □ = 1000 students

4	
199 3	□□□
199 2	□□

3. **Bar Charts:**

- Bar-diagrams are usually used to represent one way or simple frequency distribution.
- Bar-diagrams can be drawn either horizontally or vertically. Usually horizontal bar-diagrams are used for qualitatively classified data whereas vertical bar-diagrams are used for quantitatively classified data.

There are a number of bar-diagrams. The most common being:

- Simple bar chart
- Component bar chart
- Multiple bar chart

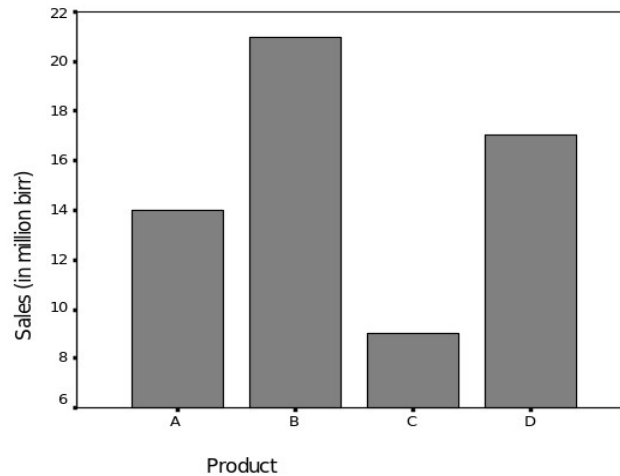
1. Simple bar-diagrams

Simple bar-diagrams are used to depict data of single variable or one-way variable.

Example: The following frequency distribution shows sales of production (in million birr) of three products for 2004 production year.

Produ ct	Sale (in million)
A	14
B	21
C	9
D	17

The bar-diagram presentation for these data is given below.



2. Component bar-diagrams

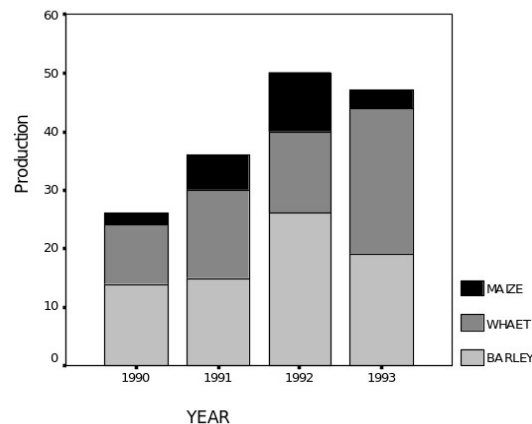
When it is desired to show how a total (an aggregate) is divided into component parts, we use component bar diagram. In such type of bar-diagrams, the bars represent aggregate value of a variable with each aggregate broken into its component parts and different colors or designs are used for identification.

Example: Represent the following data using bar-charts

Data: Yields of production of farmers in Southern Ethiopia.

Year → Crop↓	1990 EC	1991 EC	1992 EC	1993 EC
Barley	14	15	26	19
Wheat	10	15	14	25
Maize	2	6	10	3
Total	26	36	50	47

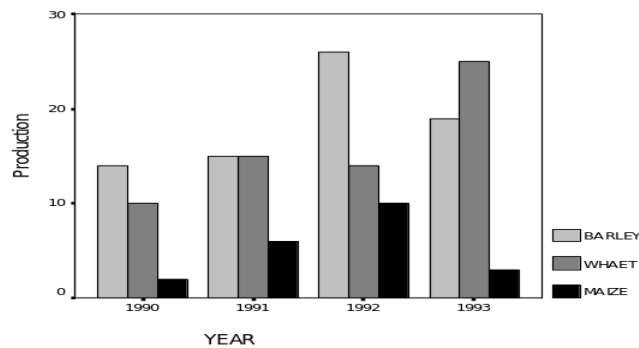
The component bar-diagram for this table is as follows



3. Multiple bar-diagrams

Multiple bar-diagrams are used to display data on more than one variable. They are used for comparing different variables at the same time.

Example: The data given in the above example can be presented by using multiple bar-diagram as below.



4. Graphical Presentation of data

Three common graphic presentations of data: histogram, frequency polygon, and cumulative frequency polygon (ogive).

Procedures for constructing statistical graphs:

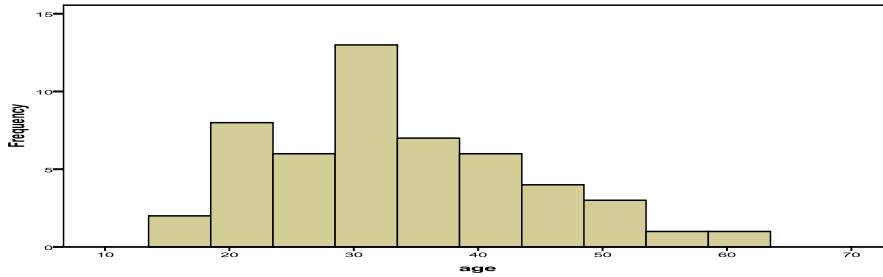
- Draw and label the X and Y axes.
- Choose a suitable scale for the frequencies or cumulative frequencies and label it on the Y axes.
- Represent the class boundaries for the histogram or ogive or the mid points for the frequency polygon on the X axes.
- Plot the points.
- Draw the bars or lines to connect the points.

a) Histogram

It presents a grouped frequency distribution of a continuous type. It is drawn by making class boundaries in the x-axis and frequencies in the y-axis.

Example: Draw a histogram for the following grouped age data.

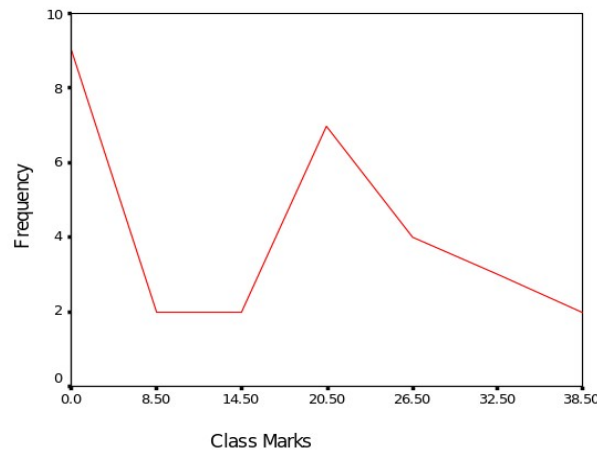
Class limit	Class boundaries	Mid point	Frequency
15-19	14.5-19.5	17	2
20-24	19.5-24.5	22	8
25-29	24.5-29.5	27	6
30-34	29.5-34.5	32	12
35-39	34.5-39.5	37	7
40-44	39.5-44.5	42	6
45-49	44.5-49.5	47	4
50-54	49.5-54.5	52	3
55-59	54.5-59.5	57	1
60-64	59.5-64.5	62	1



b. Frequency Polygon

A frequency polygon is a line graph drawn by taking the frequencies of the classes along the vertical axis and their respective class marks along the horizontal axis. Then join the cross points by a free hand curve.

Example: Present the data in the previous example using a frequency polygon.

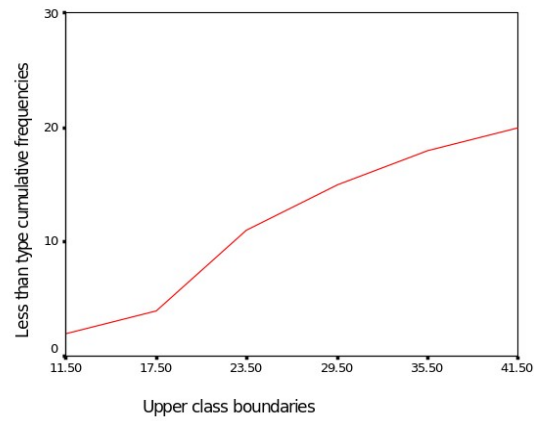


C. Cumulative Frequency Polygon (Ogive)

Cumulative frequency polygon can be traced on less than or more than cumulative frequency basis. Place the class boundaries along the horizontal axis and the corresponding cumulative frequencies (either less than or more than cumulative frequencies) along the vertical axis. Then join the cross points by a free hand curve.

Example: the data in the previous example can be presented using either a less than or a more than cumulative frequency polygon as given below (i) and (ii) respectively.

(i) Less than type cumulative frequency polygon



(ii) More than type cumulative frequency polygon

