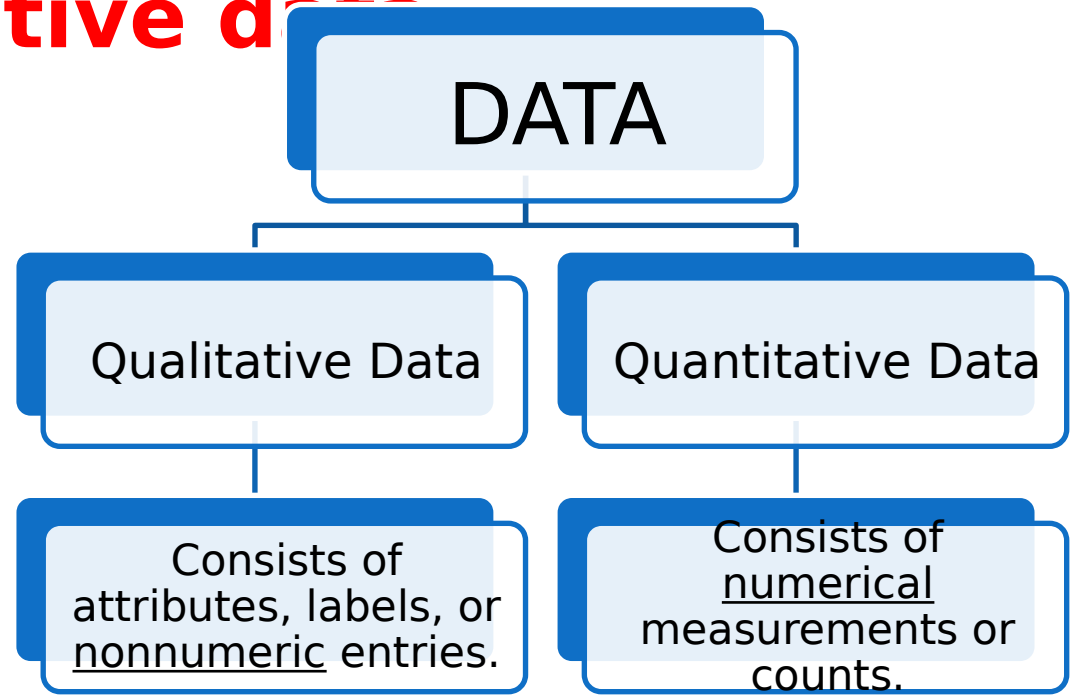
UNIT TWO

DATA COLLECTION & PRESENTATION

Types of Data

Data sets can consist of two types of data: **Qualitative data** and **Quantitative data**



Qualitative and Quantitative Data

Example: The grade point averages of five students are listed in the table. Which data are qualitative data and which are quantitative data?

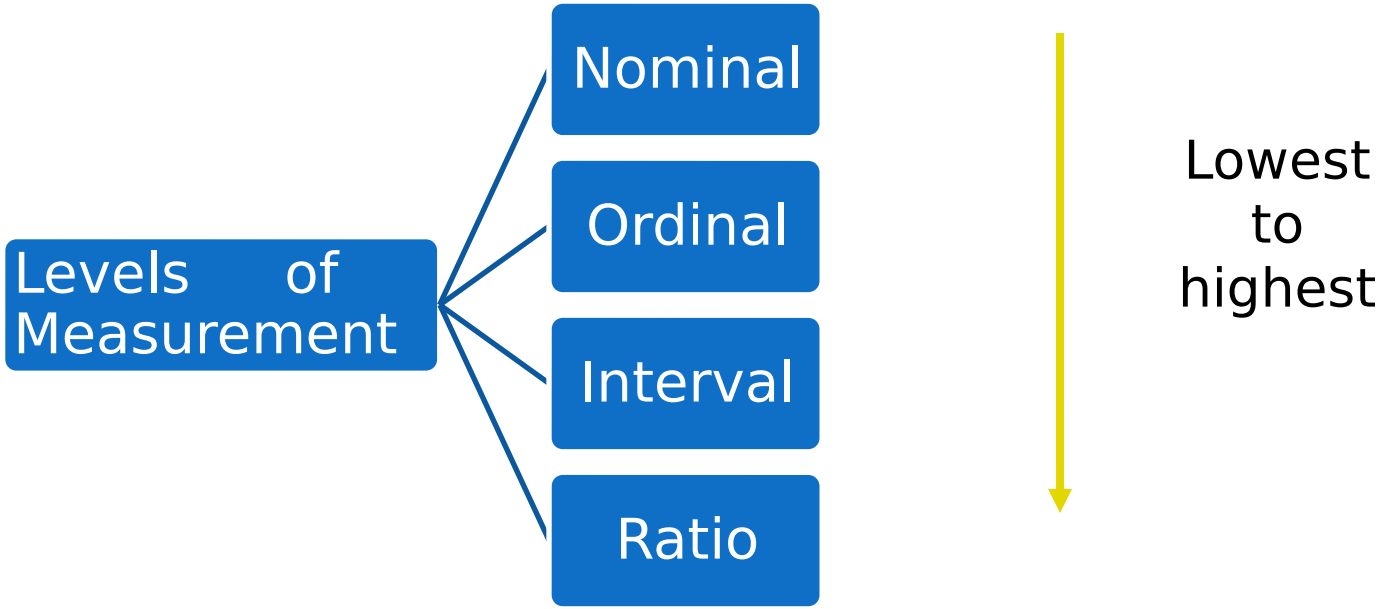
Student	GPA
Sara	3.22
Berhan	3.98
Mahlet	2.75
Tsehay	2.24
Hana	3.84

Qualitative data

Quantitative *data*

Levels of Measurement

- The level of measurement determines which statistical calculations are meaningful.
- Measurement is the assignment of values to objects or events in a systematic fashion. The four levels of measurement are: **nominal**, **ordinal**, **interval**, and **ratio**.



Nominal Scale

- The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to **distinguish** one object from another.
- Qualities with no ranking or ordering; no numerical or quantitative value. These types of data consists of names, labels and categories.
- It is a scale for grouping individuals into different categories.

  Example : ***Eye color***: brown, black, etc,

  Sex: Male, Female.
- In this scale, one is different from the other.
- Arithmetic operations (+, -, *, ÷) are not applicable, comparison (<, >, ≠, etc) is ***impossible***.

Ordinal Scale

- Defined as nominal data that **can be ordered or ranked**.
- Can be arranged in some order, but the differences between the data values are meaningless.
- Data consisting of an ordering of ranking of measurements are said to be on an ordinal scale of measurements.
- It provides enough information to order objects.
- One is different from and greater /better/ less than the other.
- Arithmetic operations (+, -, *, ÷) are impossible, comparison (<, >, ≠, etc) is ***possible***.

Example: Letter grading (A, B, C, D, F),

  - Rating scales (excellent, very good, good, fair, poor),
  - Military status (general, colonel, lieutenant, etc).

Interval Level

- Data are defined as ordinal data and the differences between data values are meaningful. However, there is no true zero, or starting point, and the ratio of data values are meaningless.
- Note: Celsius & Fahrenheit temperature readings have no meaningful zero and ratios are meaningless. For example, a temperature of zero degrees (on Celsius and Fahrenheit scales) does not mean a complete absence of heat.
- One is different, better/greater and by a certain amount of difference than another.
- Possible to add and subtract. For example; 800c – 500c = 300c, 700c – 400c = 300c.
- ***Multiplication and division are not possible***. For example; 60^0c = 3(20^0c). But this does not imply that an object which is 60^0c is three times as hot as an object which is 20^0c.
- Most common examples are: IQ, temperature.

Ratio Scale

- Similar to interval, except there is a true zero (absolute absence), or starting point, and the ratios of data values have meaning.
- Arithmetic operations (+, -, *, ÷) are applicable. For ratio variables, both differences and ratios are meaningful.
- One is different/larger /taller/ better/ less by a certain amount of difference and so much times than the other.
- This measurement scale provides better information than interval scale of measurement.
- **Example** : weight, age, number of students.

Summary of Levels of Measurement

	Levels of measurement			
	Nominal	Ordinal	Interval	Ratio
Put data in categories	Yes	Yes	Yes	Yes
Arrange data in order	No	Yes	Yes	Yes
Subtract data values	No	No	Yes	Yes
Determine if one data value is a multiple of another	No	No	No	Yes

Data Collection

- Is a ***systematic and meaningful assembly of information*** for the accomplishment of the objective of a statistical investigation.
- It refers to the methods used in gathering the required information from the units under investigation.

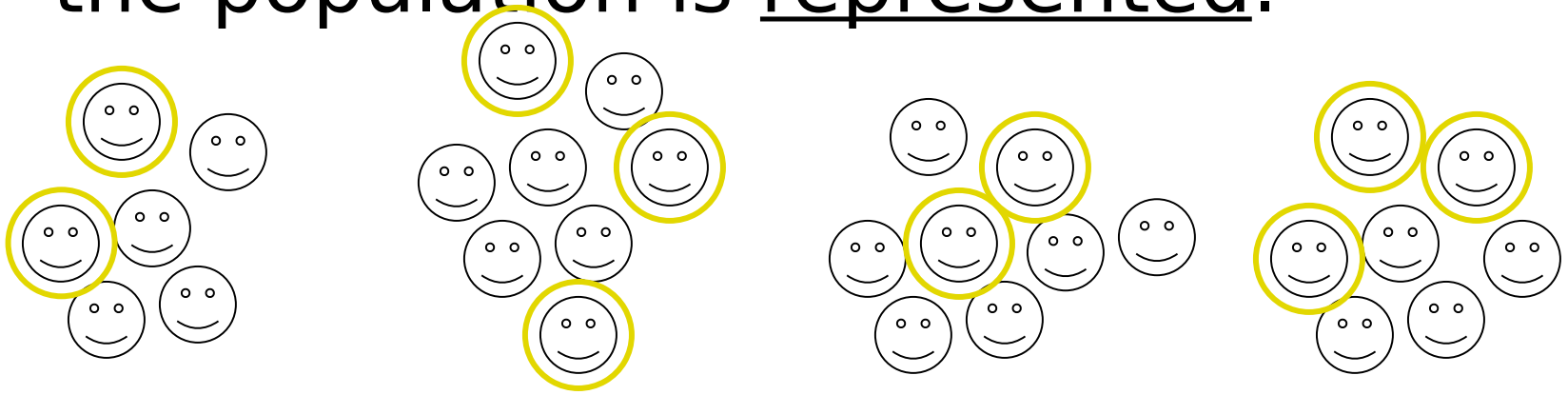
Terminologies

- A ***simulation*** is the use of a mathematical or physical model to reproduce the conditions of a situation or process.
- A ***survey*** is an investigation of one or more characteristics of a population.
- A ***census*** is a measurement of an *entire* population.
- A ***sampling*** is a measurement of *part* of a population.

Methods of Data Collection

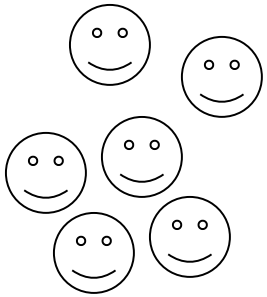
Stratified Samples

- A ***stratified sample*** has members from each segment of a population.
- This ensures that each segment from the population is <u>represented</u>.

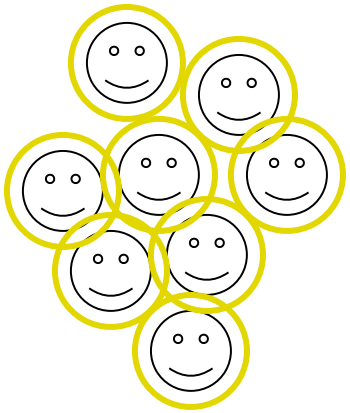


Cluster Samples

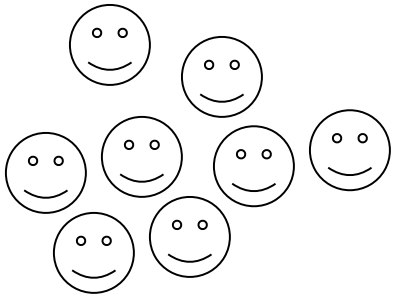
- A ***cluster sample*** has all members from randomly selected segments of a population.



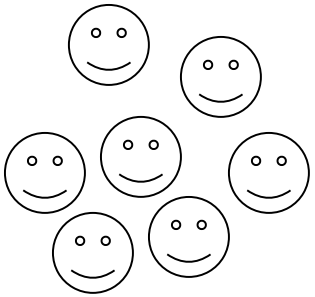
Freshmen



Sophomores



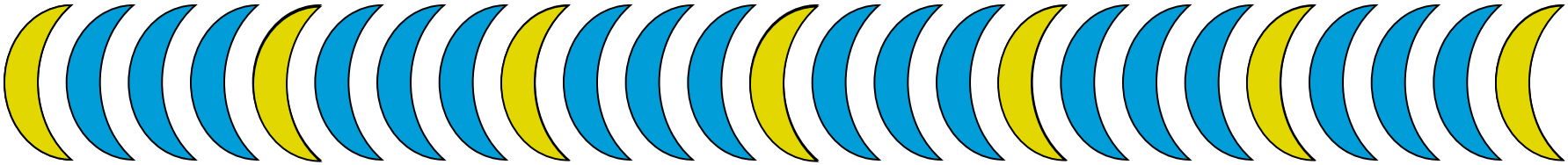
Juniors



Seniors

Systematic Samples

A ***systematic sample*** is a sample in which each member of the population is assigned a number. A starting number is randomly selected and sample members are selected at regular intervals.



Every fourth member is chosen.

Convenience Samples

- A **convenience sample** consists only of available members of the population.

•***Convenience sampling*** is sometimes referred to as haphazard or accidental sampling.

•Sample units are only selected if they can be ***accessed easily and conveniently.***

•Although useful applications of the technique are limited, it can deliver accurate results when the population is ***homogeneous.***

•May not be representative of the target population result in the presence of ***bias.***

Quota sampling

- Quota sampling
- Snowball Sampling

PRIMARY AND SECONDARY DATA

PRIMARY DATA/ SOURCES

- A primary source is a source from where first-hand information is gathered.
- Are original sources of data.

SECONDARY DATA

- Is the one that makes data available, which were collected by some other agency.
- A source, which is not primary, is necessarily a secondary source.
- Obtained from such sources as census and survey reports, books, official records, reported experimental results, previous research papers, bulletins, magazines, newspapers, web sites, and other publications.

EXAMPLE

- A study conducted to see the age distribution of HIV/AIDS victim citizens.
- Information obtained from the victim citizens are **primary sources.**
- Use of records of hospitals and other related agencies to obtain the age of the victim citizens without the need of tracing the victims personally is a **secondary source.**

Advantages and Disadvantages of Primary & Secondary data

Advantages of primary data over that of secondary data.

- Gives more reliable, accurate and adequate information, which is suitable to the objective and purpose of an investigation.
- Shows data in greater detail.
- Free from errors that may arise from copying of figures from publications, which is the case in secondary data.

DISADVANTAGES OF PRIMARY DATA

- It is time consuming and costly.
- Gives misleading information due to lack of integrity of investigators and non-cooperation of respondents.

ADVANTAGE OF SECONDARY DATA:

- It is readily available and hence convenient and much quicker
- It reduces time, cost and effort as compared to primary data.
- May be available in subjects (cases) where it is impossible to collect primary data. Such a case can be regions where there is war.

The disadvantages of Secondary data :

- Data obtained may not be sufficiently accurate.
- Data that exactly suit our purpose may not be found.
- Error may be made while copying figures.

The choice between primary data and secondary data is determined by factors

- Nature and scope of the enquiry,
- Availability of financial resources,
- Availability of time,
- Degree of accuracy desired

- Primary data are used in situations where secondary data do not provide adequate basis of analysis. i.e. when the secondary data do not suit a specific investigation.
- Unless for such cases, most statistical investigations rest up on secondary data since it minimizes cost and saves time.

Methods of collecting primary data

1. Personal Enquiry Method (Interview method)

A. Direct Personal Interview: There is a face-to-face contact with the persons from whom the information is to be obtained.

B. Indirect Personal Enquiry (Interview): The investigator contacts third parties called witnessed who are capable of supplying the necessary information.

2. Direct Observation
3. Questionnaire method

METHODS /TYPES OF CLASSIFICATION

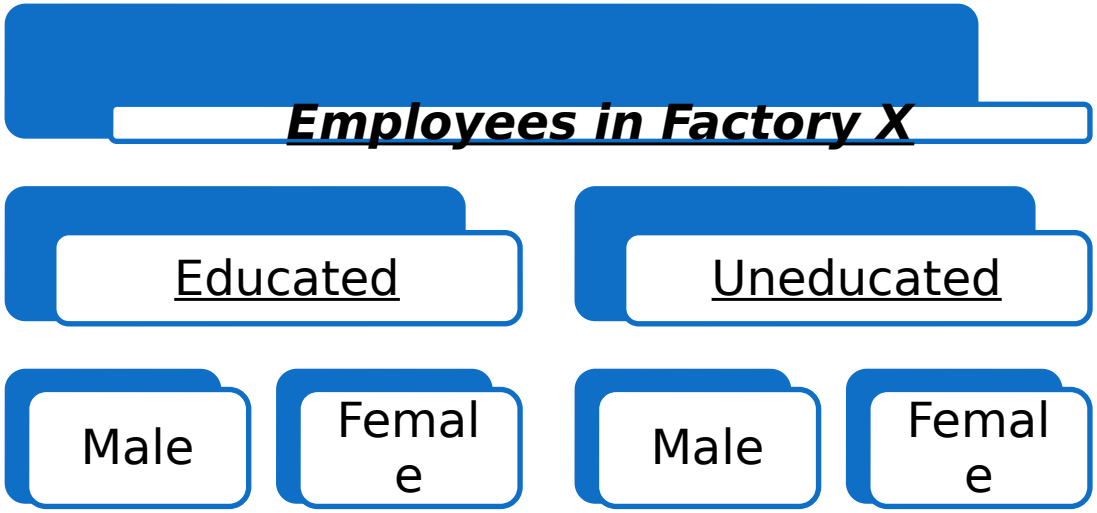
Geographical Classification: - Data are arranged according to **places** like continents, regions, and countries.

Region	Dominant Language Spoken
East Africa	Amharic
West Africa	French
North Africa	Arabic
South Africa	English

Chronological Classification:- Data are arranged according to **time** like year, month.

Year (in EC)	Population (in million)
1974	30
1986	52
1991	60

- **Qualitative Classification**: - Data are arranged according to attributes like color, religion, marital-status, sex, educational background, etc.



- **Quantitative Classification**:- The statistical data is classified according to some quantitative variables. The variable may be either discrete or continuous.

Mr. x	Height (X) in cm
A	160
B	182
C	175
D	178

- **Discrete Variables** – are variables that are associated with enumeration or counting.

Example

- Number of students in a class
- Number of children in a family, etc

- **Continuous Variables** – are variables associated with measurement.

Example

- Weights of 10 students.
- The heights of 12 persons.
- Distance covered by a car between two stations etc

FREQUENCY DISTRIBUTION

- Frequency refers to the number of observations a certain value occurred in a data.
- A frequency distribution is the organization of raw data in table form, using classes and frequencies.
- The tabular representation of values of a variable together with the corresponding frequency is called a Frequency Distribution (FD).

A. Ungrouped Frequency Distribution (UFD)

- Shows a distribution where the values of a variable are linked with the respective frequencies.
- **Example:** Consider the number of children in 15 families

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

No. of Children (Values)	No. of Family (Tallies)	Frequency
0	/ /	2
1	////	4
2	////	4
3	/ / /	3
4	/ /	2
Total		15

A. Grouped Frequency Distribution (GFD)

- If the mass of the data is very large, it is necessary to condense the data in to an appropriate number of classes or groups of values of a variable and indicate the number of observed values that fall in to each class.
- A GFD is a frequency distribution where values of a variable are linked in to groups & corresponded with the number of ...ions in each group.

Values (x_i)	1 - 25	26 - 50	51 - 75	76 - 100
Frequency	3	10	18	6

COMMON TERMINOLOGIES IN A GFD

I. **Class**:- group of values of a variable between two specified numbers called lower class limit (LCL) & upper class limit (UCL)

Class limits (CL): It separates one class from another. The limits ***could*** actually appear in the data and have ***gaps*** between the upper limits of one class and the lower limit of the next class.

In **Example*,** the GFD contains four classes:
1 – 25, 26 – 50. 51 – 75. and 76 – 100

$LCL_1 = 1$, $UCL_1 = 25$ $LCL_3 = 51$, $UCL_3 = 75$

$LCL_2 = 26$, $UCL_2 = 50$ $LCL_4 = 76$, $UCL_4 = 100$

Class boundaries: Separate one class in a grouped frequency distribution from the other. The boundary has one more decimal place than the raw data.

- There is ***no gap*** between the upper boundaries of one class and the lower boundaries of the succeeding class.
- Obtained by subtracting half of the unit of measurement (u) from the lower limits and by adding ½ (u) on the upper limits of a class. U can assume values 1, 0.1, 0.01, 0.001……

i.e $UCB_i = UCL_i + ½ (u)$

$LCB_i = LCL_i - ½ (u)$

Where UCB_i = Upper Class Boundaries and

ii. <u>Class Frequency (or Simply Frequency)</u>: refers to the number of observations corresponding to a class.

In **Example** * The class frequency of the 1^{st}, 2^{nd}, 3^{rd}, & 4^{th} classes are respectively 3, 10, 18 and 6.

Note: The unit of measurement (u) is the gap between any two successive classes. i.e

u = lower limit of a class – upper limit of the preceding class.

In **Example *,** consider the 2nd class, 26 – 50, since u = 26 – 25 = 1,

$LCL_2 = 26$ $UCL_2 = 50$

$LCB_2 = 26 - ½(1) = 25.5$ $UCB_2 = 50 + ½(1) = 50.5$

iv. Class Width (size of a class or class interval): it is the difference between the upper and lower class limits or the difference between the upper and lower class boundaries of any class.

Remarks:

1. If both the LCL & UCL are included in a class, it is called an inclusive class. For inclusive classes,

   Class width (cw) = UCB_i - LCB_i

2. If LCL is included and the UCL is not included in a class, it is called an exclusive class. For exclusive classes;

   Class width (cw) = UCL_i – LCL_i

To be consistent, we use inclusive classes

v. **Class Mark (cm):** it is the mid point (center) of a class

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

Note:- the difference between any two successive class marks is equal to the width of a class

Range (R) : is the difference between the largest (L) and the smallest (S) values in a data

$$R = L - S$$

RULES FOR FORMING A GROUPED FREQUENCY DISTRIBUTION

To construct a GFD the following points should be considered

1. The classes should be clearly defined. That is each observation should fall in to one & only one class.
2. The number of classes neither should be too large nor too small. Normally, 5 to 20 classes are recommended.
3. All the classes should be of the same width. An approximate suitable class width can be obtained as:

$$cw \approx \frac{Range}{Number\ of\ Classes} \quad i.e \quad cw = \frac{R}{n} = \frac{L - S}{n}$$

Note that a suitable number of classes can be obtained by using the formula

$$n \approx 1 + 3.322 \log N.$$

up/down to the nearest whole number, where N is the total number of observations.

- Alternatively n can also be determined by formula

Where

n=Number of Classes

N=Total number of observations

4. Determine the class limits

- Determine the lower class limit of the first class (LCL1), then

- $LCL_2 = LCL_1 + cw$, $LCL_3 = LCL_2 + cw$,... $LCL_{i+1} = LCL_i + cw$
  - Determine the upper class limit of the first class (UCL_1) i.e.

    $UCL_1 = LCL_1 + cw - u$,
  - where u = the unit of measurement, then

    $UCL_2 = UCL_1 + cw$, UCL_3 UCL_2, ... , $UCL_{i+1} = UCL_i + cw$
  - Complete the GFD with the respective class frequencies.

- **Example.** The number of customers for consecutive 30 days in a supermarket was listed as follows:

20	48	65	25	48	49
35	25	72	42	22	58
53	42	23	57	65	37
18	65	37	16	39	42
49	68	69	63	29	67

A. Construct a GFD with a suitable number of classes
B. Complete the distribution obtained in (A) with class boundaries & class marks

Solution: i. Range = Largest value – smallest value

= 72 – 16 = 56

N = 30 (total number of observations)

$\therefore$ number of classes, n = 1 + 3.322 log30

$\Rightarrow$ n = 1 + 3.322 log30

= 1 + 3.322 (1.4771)

= 5.9

- Hence a suitable number of class n is chosen to be 6

$$cw \approx \frac{Range}{Number\ of\ Classes} \quad i.e \quad cw = \frac{R}{n} = \frac{L - S}{n}$$

- Class width = 9.33 = **cw**
- For the sake of convenience, take **cw** to be 10 (note that it is also possible to choose the **cw** to be 9).
- Take lower limit of the 1st class (LCL_1) to be 16 & u = 1
- i.e. $LCL_1 = 16$ and $UCL_1 = LCL_1 + cw - u = 16+10-1 = 25$

  $LCL_2 = LCL_1 + cw = 16 + 10 = 26$ $UCL_2 = UCL_1 + cw = 25 + 10 = 35$

  $LCL_3 = LCL_2 + cw = 26 + 10 = 36$ $UCL_3 = UCL_2 + cw = 35 + 10 = 45$
- Therefore, the GFD would be

A)

Class (x_i)	Frequency (f_i)
16 – 25	7
26 – 35	2
36 – 45	6
46 – 55	5
56 – 65	6
66 – 75	4

B)

Class (x_i)	Frequency (f_i)	CB_i	cm_i
16 – 25	7	15.5 – 25.5	20.5
26 – 35	2	25.5 – 35.5	30.5
36 – 45	6	35.5 – 45.5	40.5
46 – 55	5	45.5 – 55.5	50.5
56 – 65	6	55.5 – 65.5	60.5
66 – 75	4	65.5 – 75.5	70.5

CUMULATIVE FREQUENCY DISTRIBUTION (CFD)

- **Cumulative frequency (CF):** It is the number of observation less than the upper class boundary or greater than the lower class boundary of class.
- **‘Less Than’ Cumulative Frequency Distribution (<CFD):** it is the number of values less than the upper class boundary of a given class.
- **‘More Than’ Cumulative Frequency Distribution (>CFD):** it is the number of values greater than

Example : Consider the frequency distribution given below

Class (x_i)	Frequency (f_i)	Less than Cumulative Frequency ($<cf_i$)	More than Cumulative Frequency ($>cf_i$)
3 – 6	4	4	30
7 – 10	7	11	26
11 – 14	10	21	19
15 – 18	6	27	9
19 – 22	3	30	3

This means that from ‘less than’ cumulative frequency distribution there are 4 observations less than 6.5, 11 observations below 10.5, etc and from ‘more than’ cumulative frequency distribution 30 observations are above 2.5, 26 above 6.5 etc.

RELATIVE FREQUENCY DISTRIBUTION (RFD)

- It enables the researcher to know the proportion or percentage of cases in each class.
- Obtained by dividing the frequency of each class by the total frequency. It can be converted in to a percentage frequency by multiplying each relative frequency by 100%. i.e.

$$Rf_i = \frac{f_i}{n}$$

- Where Rfi – is the relative frequency of the ith class

f_i – is the frequency of the i^{th} class

n – is the total number of observations

Note: $Pf_i = Rf_i \times 100\%$

Example : The relative and percentage frequency distribution of is :

x_i	f_i	Rf_i		%freq. (Pf_i)
3 – 6	4	4/30	0.13	$4/30 \times 100$
7 – 10	7	7/30	0.23	$7/30 \times 100$
11 – 14	10	10/30	0.33	$10/30 \times 100$
15 – 18	6	6/30	0.20	$6/30 \times 100$
19 – 22	3	3/30	0.10	$3/30 \times 100$
Total	30	1	100%	100%

Relative cumulative frequency (RCf): The running total of the relative frequencies or the cumulative frequency divided by the total frequency gives the percent of the values which are less than the upper class boundary or the reverse.

$$CRf_i = Cf_i/n = Cf_i/\sum f_i$$

PRESENTATION OF DATA

- **Presentation** is a statistical procedure of arranging and putting data in a form of tables, graphs, charts and/or diagrams.

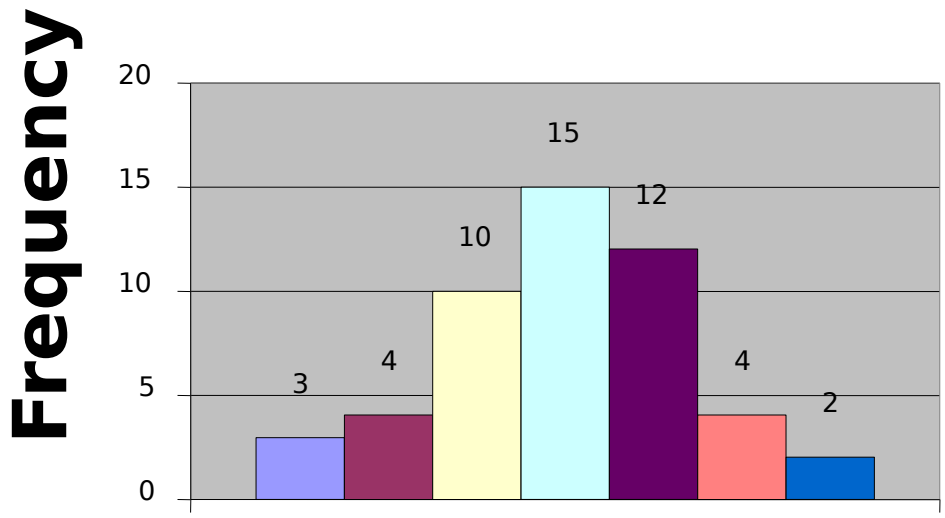
HISTOGRAM

- Consisting of a series of adjacent rectangles whose bases are equal to the class width of the corresponding classes and whose heights are proportional to the corresponding class frequencies.
- The class boundaries are marked along the x – axis and the class frequencies along the y – axis.
- It describes the shape (symmetry) of the data and where do most of the data values lie?

- **Example :** A histogram to representing the following data.

Class limits	15-24	25-34	35-44	45-54	55-64	65-74	75-84
Frequency	3	4	10	15	12	4	2

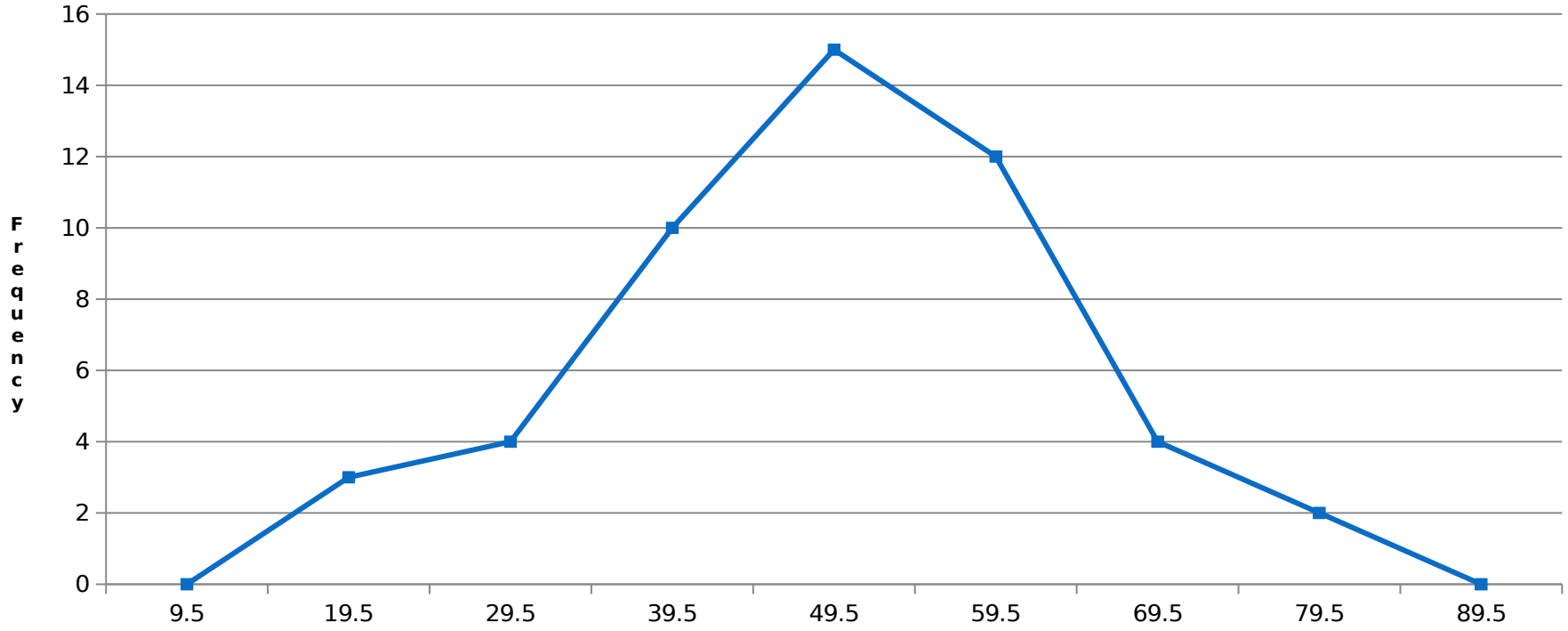
Histogram



FREQUENCY POLYGON

- It is a line graph of frequency distribution.
- Clearly illustrates shape of the data than a histogram does.
- Connects ***the centers*** (class marks) of the tops of the histogram bars with a series of straight lines.

Frequency Polygon

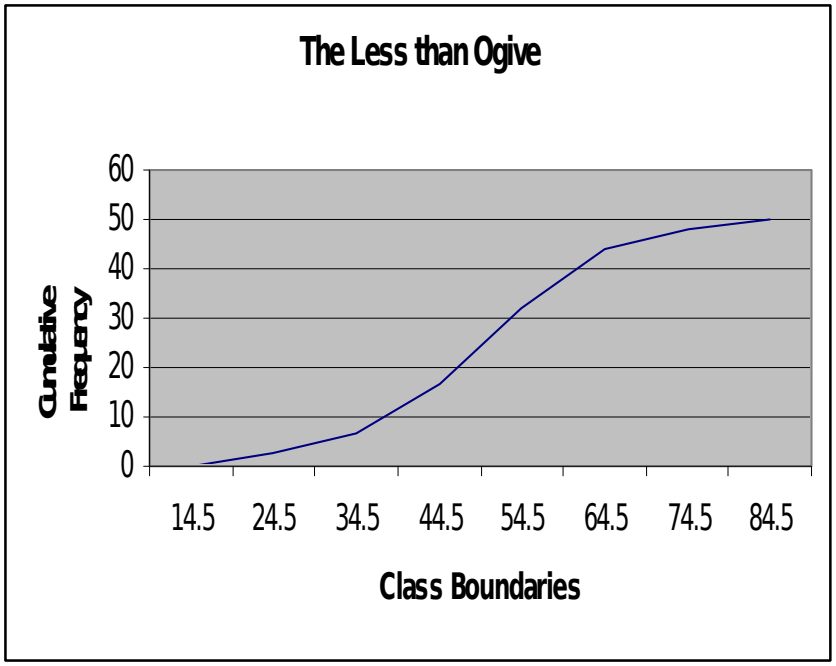


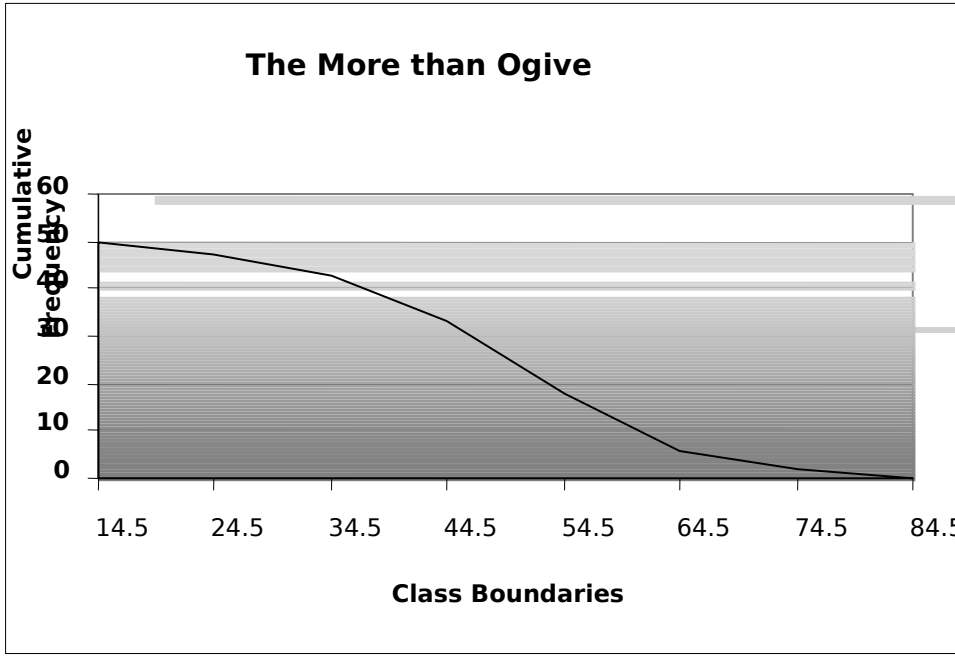
Class mark

CUMULATIVE FREQUENCY CURVE, (OGIVE)

- *It is useful for determining the number of values below or above some particular value.*
- *Uses class boundaries along the horizontal axis and frequencies along the vertical axis.*
- *There are two type of O-give namely less than Ogive and more than Ogive.*

CUMULATIVE FREQUENCY CURVE, (OGIVE)

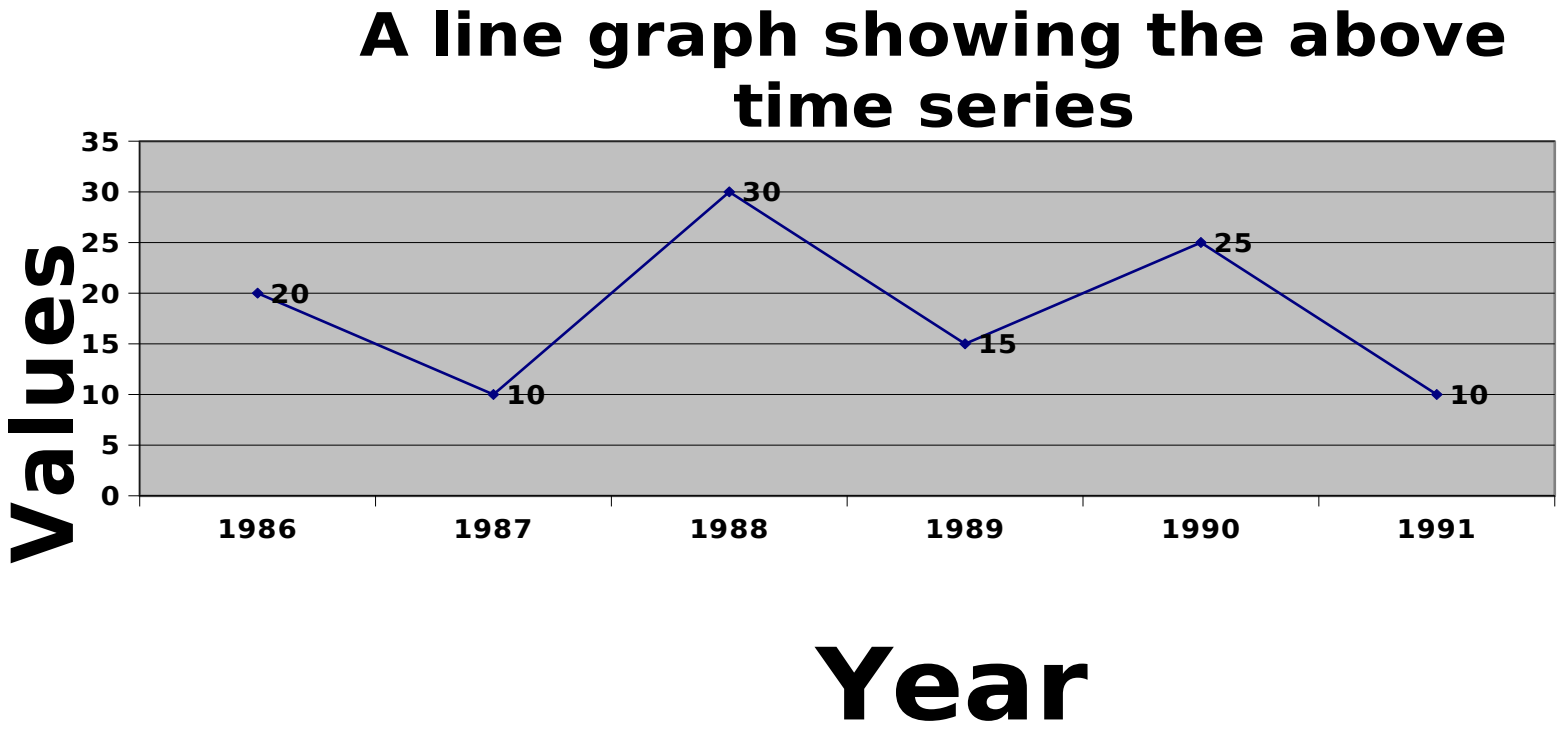




LINE GRAPH

Example . Draw a line graph for the following time series.

Year	1986	1987	1988	1989	1991
Values	20	10	30	15	1

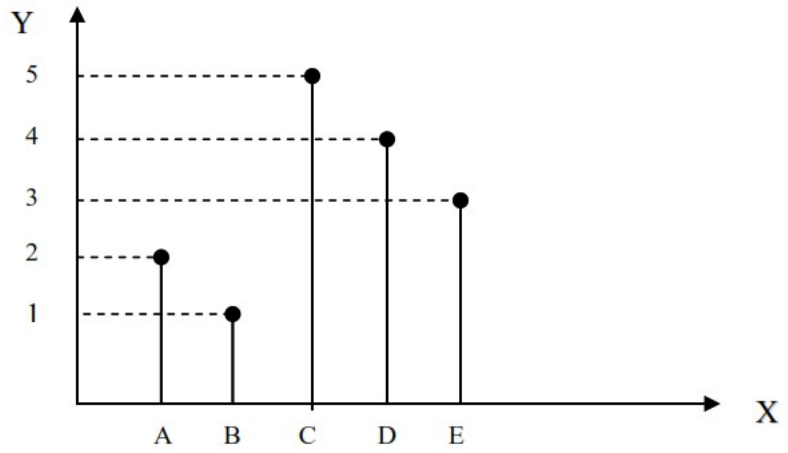


VERTICAL LINE GRAPH

- Is a graphical representation of discrete data and frequencies.
- Vertical solid lines are used to indicate the frequencies.
- **Example .** Draw a vertical line graph for the following data

Family	A	B	C	D	E
Number of children	2	1	5	4	3

Solution:



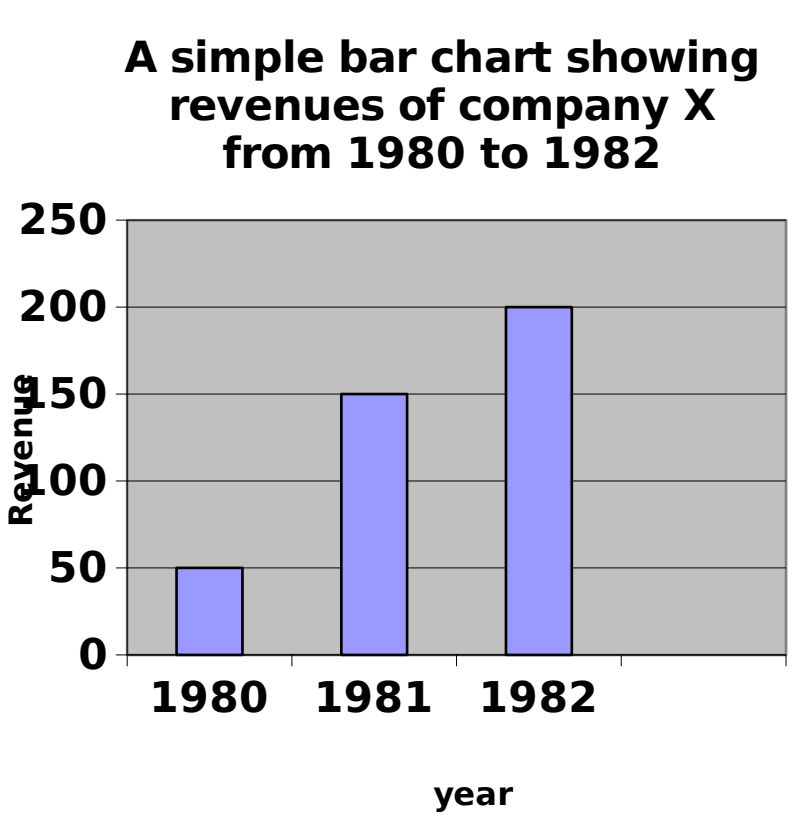
Vertical line graph showing number of children in family A, B, C, D and E

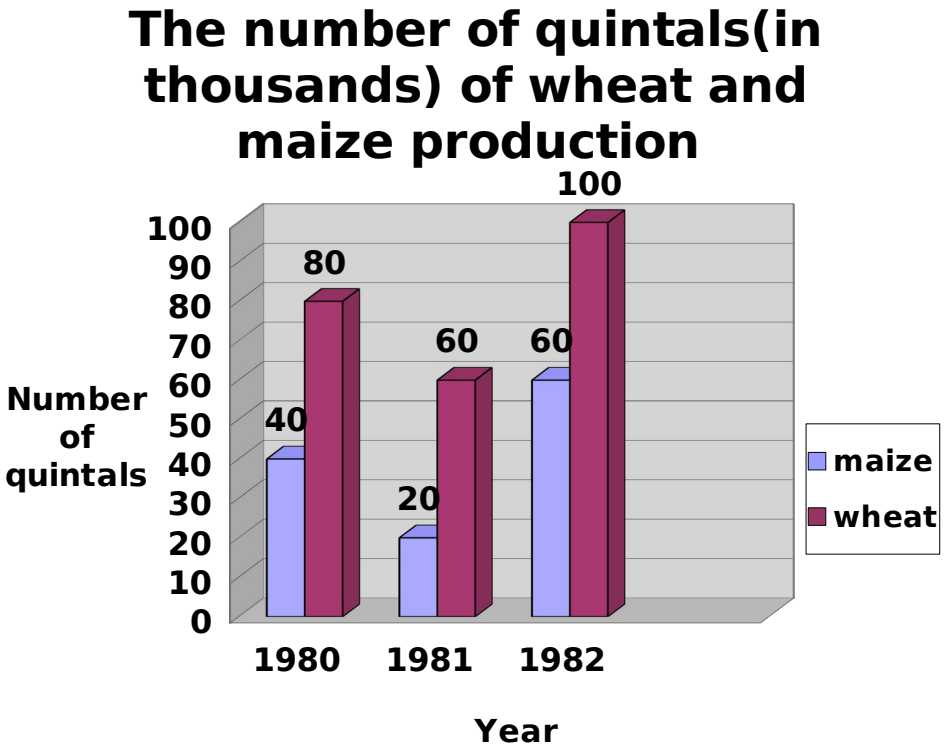
BAR CHART (BAR DIAGRAM)

- Histogram, Frequency polygon, ogives are used for data having an interval or ratio level of measurement.
- Bar chart is a series of equally spaced bars of uniform width where the height (length) of a bar represents the frequency corresponding with a category.
- Bars may be drawn horizontally or vertically. Vertical bar graphs are preferred as they allow comparison with other bars.
- **Example:** Revenue (in millions of Birr) of company x from 1980 to 1982 is given below

Year	Revenue
1980	50
1981	150
1982	200

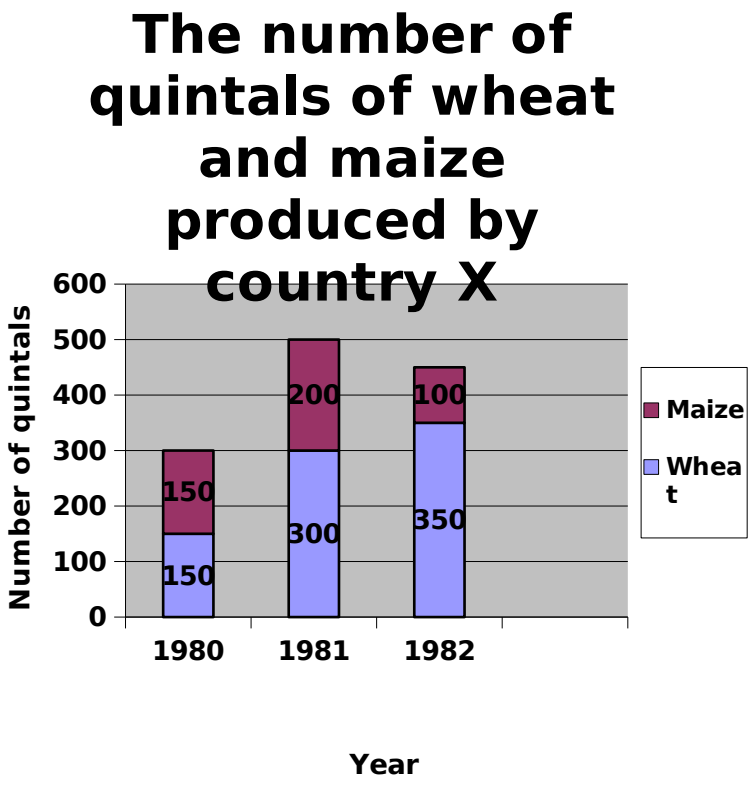
Year	Maize	Wheat
1980	40	80
1981	20	60
1982	60	100





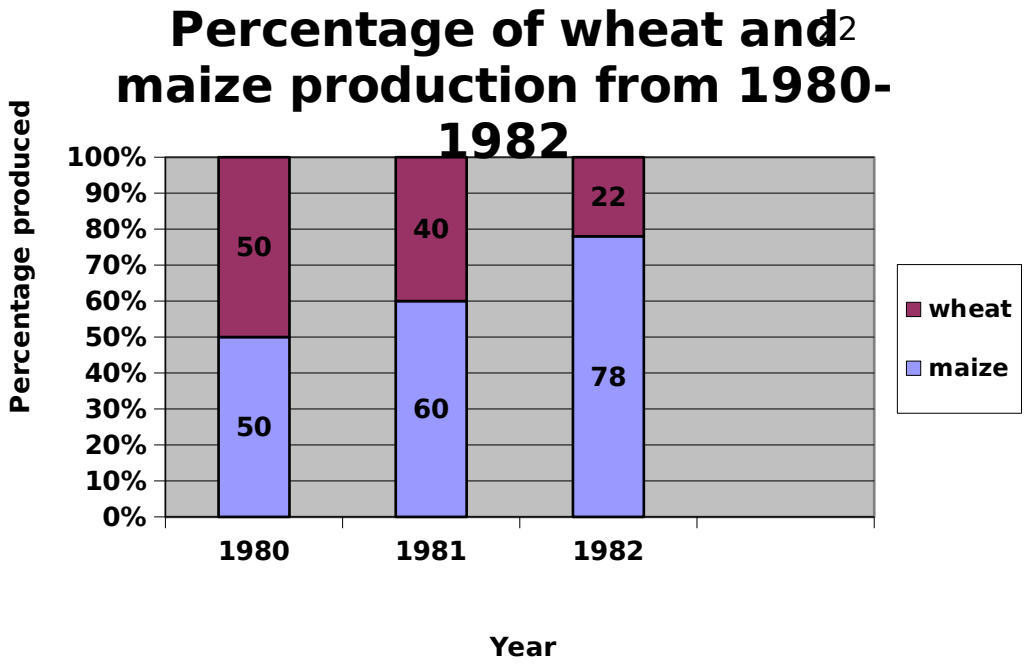
SUBDIVIDED BAR CHART

Year	Wheat	Maize
1980	150	150
1981	300	200
1982	350	100



Example : percentage bar chart

Year	% of Wheat Production	% of Maize Production
1980	150/300 × 100 = 50	150/300 × 100 = 50
1981	300/500 × 100 = 60	200/500 × 100 = 40
1982	350/450 × 100 = 78	100/450 × 100 = 22

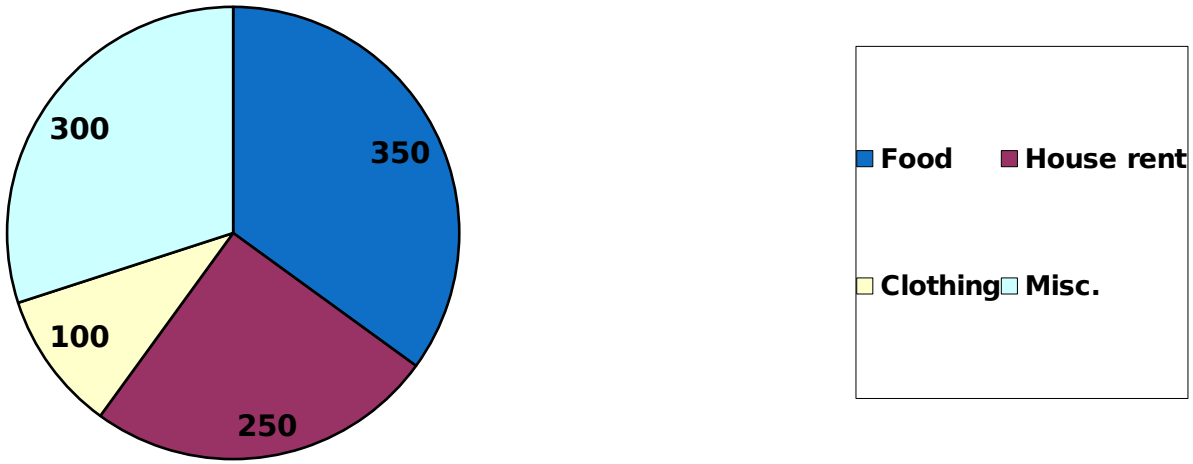


PIE CHART

- A ***pie chart*** is a circle that is divided in to sections or according to the percentage of frequencies in each category of the distribution.
- **Example:** The monthly expenditure of a certain family is given below.

Items	Expenditure	% Proportion (Pf_i)	Degrees (360^o Rf_i)
Clothing	100	$100/1000 \times 100 = 10$	$100/1000 \times 360^o = 36$
Food	350	$350/1000 \times 100 = 35$	$350/1000 \times 360^o = 126$
House Rent	250	$250/1000 \times 100 = 25$	$250/1000 \times 360^o = 90$
Miscellaneous	300	$300/1000 \times 100 = 30$	$300/1000 \times 360^o = 108$
Total	1000	100%	360^o

Solution: The pie chart for the above expenditure is as follows



PICTOGRAPH (PICTOGRAM)

- A pictograph is a graph that uses symbols or pictures to represent data.
- **Example :** In comparing the population of a country from 1990 to 1992, we simply draw pictures of people where each picture may represent 1000,000 people.

1992 - □□□□□□□ Key: □ = 1,000,000

1991 - □□□□□□□□□

1990 - □□□□□□□□□□□□□□□


Thanks!