Introduction to Statistics for Engineers

Chapter One

Introduction to Statistics

Outline

- Introduction to statistics
- Importance of statistics
- Limitation of statistics
- Application areas of statistics
- Classification of statistics
- Main terms in statistics
- Steps In Statistical Investigation
- The Engineering Method And Statistical Thinking

Definitions:

- It is a science which helps us to collect, analyze and present data systematically.
- It is the process of collecting, processing, summarizing, presenting, analysing and interpreting of data in order to study and describe a given problem.
- Statistics is the art of learning from data.
- Statistics may be regard as (i)the study of populations, (ii) the study of variation, and (iii) the study of methods of the reduction of data.

Importance of Statistics:

- It simplifies mass of data (condensation);
- Helps to get **concrete** information about any problem;
- Helps for reliable and objective decision making;
- It presents facts in a precise & definite form;
- It facilitates comparison(Measures of central tendency and measures of dispersion);
- It facilitates Predictions (Time series and regression analysis are the most commonly used methods towards prediction.)
- It helps in formulation of suitable policies;

Limitation of statistics:

- Statistics does not deal with **individual items;**
- Statistics deals only with **quantitatively expressed items**, it does not study qualitative phenomena;
- Statistical results are **not universally true**;
- Statistics is **liable to be misused**.

Application areas of statistics

Some of the diverse fields in which Statistical methodology has extensive applications are:

- **Engineering:**

  Improving product design, testing product performance, determining reliability and maintainability, working out safer systems of flight control for airports, etc.

- **Business:**

  Estimating the volume of retail sales, designing optimum inventory control system, producing auditing and accounting procedures, improving working conditions in industrial plants, assessing the market for new products.

Cont’d

- **Quality Control:**

  Determining techniques for evaluation of quality through adequate sampling, in process control, consumer survey and experimental design in product development etc.

  * Realizing its importance, large organizations are maintaining their own Statistical Quality Control Department *.

- **Economics:**

  Measuring indicators such as volume of trade, size of labor force, and standard of living, analyzing consumer behavior, computation of national income accounts, formulation of economic laws, etc.

  * Particularly, Regression analysis extensively uses in the field of Economics*.

Cont’d

- **Health and Medicine:**

  Developing and testing new drugs, delivering improved medical care, preventing diagnosing, and treating disease, etc. Specifically, inferential Statistics has a tremendous application in the fields of health and medicine.

- **Biology:**

  Exploring the interactions of species with their environment, creating theoretical models of the nervous system, studying genetically evolution, etc.

- **Psychology:**

  Measuring learning ability, intelligence, and personality characteristics, creating psychological scales and abnormal behavior, etc.

- **Sociology:**

  Testing theories about social systems, designing and conducting sample surveys to study social attitudes, exploring cross-cultural differences, studying the growth of human population, etc.

Classification of statistics

There are two main branches of statistics:

1. Descriptive statistics
2. Inferential statistics

1. Descriptive statistics:

- It is the **first phase** of Statistics;
- involve any kind of data processing designed to the collection, organization, presentation, and analyzing the important features of the data **with out attempting to infer/conclude any thing that goes beyond the known data.**
- In short, descriptive Statistics describes the nature or characteristics of the observed data (usually a sample) **without making conclusion or generalization.**

The following are some examples of descriptive Statistics:

- The daily average temperature range of AA was 25 0c last week .
- The maximum amount of coffee export of Eth. (as observed from the last 20 years) was in the year 2004.
- The average age of athletes participated in London Marathon was 25 years.
- 75% of the instructors in AAU are male.
- The scores of 50 students in a Mathematics exam are found to range from 20 to 90.

2. Inferential statistics (Inductive Statistics)

- It is a second phase of Statistics which deals with techniques of making a generalization that lie outside the scope of Descriptive Statistics;
- It is concerned with the process of drawing conclusions (inferences) about specific characteristics of a population based on information obtained from samples;
- It is a process of performing hypothesis testing, determining relationships among variables, and making predictions.
- The area of inferential statistics entirely needs the whole aims to give reasonable estimates of unknown population parameters.

Cont’d

The following are some examples of inferential Statistics:

- The result obtained from the analysis of the income of 1000 randomly selected citizens in Ethiopia suggests that the average monthly income of a citizen is estimated to be 600 Birr.
- Here in the above example we are trying to represent the income of about entire population of Ethiopia by a sample of 1000 citizens, hence we are making inference or generalization.
- Based on the trend analysis on the past observations/data, the average exchange rate for a dollar is expected to be 21 birr in the coming month.
- The national statistical Bereau of Ethiopia declares the out come of its survey as “The population of Eth. in the year 2020 will likely to be 120,000,000.”
- From the survey obtained on 15 randomly selected towns of Eth. it is estimated that 0.1% of the whole urban dwellers are victims of AIDS virus.

Exercise: Descriptive Statistics or Inferential Statistics

1. The manager of quality control declares the out come of its survey as “the average life span of the imported light bulbs is 3000 hrs.
   Inferential
2. Of all the patients taking the drug at a local health center 80% of them suffer from side effect developed.
   Descriptive
4. The national statistical bureau of Eth. declares the out come of its survey for the last 30 years as “the average annual growth of the people of Ethiopia is 2.8%. Inferential
5. Based on the survey made for the last 10 years 30,000 tourists are expected to visit Ethiopia. Descriptive.
6. Based on the survey made for the last 20 years the maximum number of tourists visited Eth. were in the year 1993. Descriptive.
7. The Ethiopian tourism commission has announced that (as observed for the last 20 years) the average number of tourists arrived Ethiopia per year is 3000. Inferential
8. The maximum difference of the salaries of the workers of the company until the end of last year was birr 5000. Descriptive.

Main terms in statistics

Data: Certainly known facts from which conclusions may be drawn.

Statistical data: Raw material for a statistical investigation which are obtained when ever measurement or observations are made.

i. **Quantitative data:** data of a certain group of individuals which is expressed numerically.

Example: Heights, Weights, Ages and, etc of a certain group of individuals.

ii. **Qualitative data:** data of a certain group of individuals that is not expressed numerically as it is. Example: Colors, Languages, Nationalities, Religions, health, poverty etc of a certain group of individuals.

Cont'd

Variable: A variable is a factor or characteristic that can take on different possible values or outcomes. Example: Income, height, weight, sex, age, etc of a certain group of individuals are examples of variables. A variable can be qualitative or quantitative (numeric).

Population: A complete set of observation (data) of the entire group of individuals under consideration e.g. The number of students in this class, the population in Addis Ababa etc. A population can be finite or infinite.

Sample: A set of data drawn from population containing a part which can reasonably serve as a basis for valid generalization about the population.Or a sample is a portion of a population selected for further analysis.

Sample size: The number of items under investigation.

Cont’d

Survey (experiment): it is a device of obtaining the desired data. Two types of survey:

1. **Census Survey:** A way of obtaining data referring the entire population which is said to provide a total coverage of the population.
2. **Sample Survey:** A way of obtaining data referring a portion of the entire population which is said to provide only a partial coverage of the population.

Steps/stages in Statistical Investigation

1. Collection of Data:

Data collection is the process of gathering information or data about the variable of interest. Data are inputs for Statistical investigation. Data may be obtained either from primary source or secondary source.

2. Organization of Data

Organization of data includes three major steps.

1. *Editing:* checking and omitting inconsistencies, irrelevancies.
2. *Classification :* task of grouping the collected and edited data .
3. *Tabulation:* put the classified data in the form of table.

Cont'd

3. Presentation of Data

The purpose of presentation in the statistical analysis is to display what is contained in the data in the form of Charts, Pictures, Diagrams and Graphs for an easy and better understanding of the data.

4. Analyzing of Data

- In a statistical investigation, the process of analyzing data includes finding the various statistical constants from the collected mass of data such as measures of central tendencies (averages) , measures of dispersions and soon.
- It merely involves mathematical operations: different measures of central tendencies (averages), measures of variations, regression analysis etc. In its extreme case, analysis requires the knowledge of advanced mathematics.

Cont’d

5. Interpretation of Data

- involve interpreting the statistical constants computed in analyzing data for the formation of valid conclusions and inferences.
- It is the most difficult and skill requiring stage.
- It is at this stage that Statistics seems to be very much viable to be misused.
- Correct interpretation of results will lead to a valid conclusion of the study and hence can aid in taking correct decisions.
- Improper (incorrect) interpretation may lead to wrong conclusions and makes the whole objective of the study useless.

19

LECTURE TWO

THE ENGINEERING METHOD AND STATISTICAL THINKING

- An engineer is someone who solves problems of interest to society by the efficient application of scientific principles.

- Engineers accomplish this by either refining an existing product or process or by designing a new product or process that meets customers' expectations and needs.

The steps in the engineering method are as follows:

Develop a clear and brief description of the problem.

1. Identify, the important factors that affect this problem or that may play a role in its solution.
2. Propose a model for the problem, using scientific or engineering knowledge of the phenomenon being studied. State any limitations or assumptions of the model.
3. Conduct appropriate experiments and collect data to test or validate the tentative model or conclusions made in steps 2 and 3.

Cont'd

5. Refine the model on the basis of the observed data.
6. Manipulate the model to assist in developing a solution to the problem.
7. Conduct an appropriate experiment to **confirm** that the proposed solution to the problem is both effective and efficient.
8. Draw conclusions or make recommendations based on the problem solution.

Cont’d

- The engineering method features a strong interplay between the problem, the factors that may influence its solution, a model of the phenomenon, and experimentation to verify the adequacy of the model and the proposed solution to the problem.
- Specifically, statistical techniques can be a powerful aid in designing new products and systems, improving existing designs, and designing, developing, and improving production processes.

Cont’d

- Therefore, Engineers must know how to efficiently plan experiments, collect data, analyze and interpret the data, and understand how the observed data are related to the model that have been proposed for the problem under study.

Data collection and representation

- The term “Data Collection” refers to all the issues related to data sources, scope of investigation and sampling techniques.

- The quality of data greatly affects the final output of an investigation.

- With inaccurate and inadequate data, the whole analysis is likely to be faulty and also the decisions to be taken will also be misleading.

Cont'd

Classification of Data based on source:

1. **Primary**: data collected for the purpose of specific study.

   It can be obtained by:

   - Direct personal observation
   - Direct or indirect oral interviews
   - Administrating questionnaires

2. **Secondary**: refers to data collected earlier for some purpose other than the analysis currently being undertaken.

   It can be obtained from:

   - External Secondary data Sources(for eg. gov't and non gov't publications)
   - Internal Secondary data Sources: the data generated within the organization in the process of routine business activities

Advantages and Disadvantages of Primary and Secondary data

The advantages of primary data over of secondary data:-

- The primary data gives more reliable, accurate and adequate information, which is suitable to the objective and purpose of an investigation.
- Primary source usually shows data in greater detail.
- Primary data is free from errors that may arise from copying of figures from publications.

Cont’d

The disadvantages of primary data are:

- The process of collecting primary data is time consuming and costly.
- Often, primary data gives misleading information due to lack of integrity of investigators and non-cooperation of respondents in providing answers to certain delicate questions.

The advantage of Secondary data:

Cont'd

- It is readily available and hence convenient and much quicker to obtain than primary data.
- It reduces time, cost and effort as compared to primary data.
- Secondary 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 are:

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

Cont'd

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

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

Cont'd

Methods of collecting primary data

I. Personal Enquiry Method (Interview method)

II. Direct Observation (personal observation)

III. Questionnaire method (written questions)

I. Personal Enquiry Method (Interview method):

- In personal enquiry method, a question sheet is prepared which is called schedule.
- Depending on the nature of the interview, personal enquiry method is further classified into two types.

1. Direct Personal Interview
2. Indirect Personal Enquiry (Interview)

Cont'd

II. Direct Observation:

- In this approach, an investigator stays at the place of survey and notes down the observation himself.

III. Questionnaire Method:

- Under this method, a list of questions related to the survey is prepared and sent to the various respondents by post, Web sites, e-mail, etc.

Sampling Techniques

- Inferential statistics is a systematic method of inferring satisfactory conclusions about the population on the basis of examining a few representative units termed as sample.
- The process of selecting samples is called sampling
- The number of units in the sample is called Sample size.
- The size of sample for a study is determined on the basis of the following factors
- The size of the population
- The availability of resources
- The degree of accuracy
- The homogeneity or heterogeneity of the population
- The nature of the study
- The method of sampling technique adopted
- The nature of respondents

Reason for sampling

We study a sample of population instead of considering the entire population due to one or the other of the following reasons.

- Shortage of money, time and labor
- Limited data available
- Minimize destruction
- Obtaining more complete detailed and accurate information about the characteristics of the population with less time, energy and expenditure

Cont'd

A good sample possesses two characteristics, which are:

i. **Representativeness of the Universe and**

ii. **Adequate in Size**

- So the selection of a sample should be done in a manner that every item in the universe must have an equal chance of inclusion in the sample.

The sampling methods may broadly be classified as:

i. Random/Probability sampling

ii. Non random/Non-probability sampling

Cont'd

- Random sampling method is a method of selection of a sample such that each item within the population has equal chance of being selected.

Random sampling method is further divided into the following

i. Simple

ii. Stratified,

iii. Systematic and

iv. Cluster sampling

i. Simple Random Sampling Method:

This method involves very simple method of drawing a sample from a given population used when the population under consideration is homogenous.

Cont'd

ii. Stratified Random Sampling Method:

This method of sampling is used when the population under study is heterogeneous.

iii. Systematic Random Sampling Method:

This method of sampling is a common method of selecting a sample when a complete list of the population is available

iv. Cluster Random Sampling:

In cluster random sampling, the total population is divided into a number of relatively small non overlapping divisions called clusters and some of these clusters can randomly be selected for the inclusion of the over all sample.

Non-random (Non-probability) Sampling Method

non-random sampling technique is further divided into the following:

- Judgment
- Convenient and
- Quota sampling.

Judgment Sampling: -

In Judgment sampling, personal judgment plays a significant role in the selection of the sample

Convenient Sampling: - Convenient Sample is a Sample obtained by selecting population units that are convenient to select for the investigator with regard to saving time, money and labor and obtaining required information.

Cont'd

Quota sampling: - In Quota sampling the population is subdivided into a number of strata's (groups) and instead of taking random samples from each stratum, investigator takes simply quotas from the different strata.

The sample with in prescribed quota is selected by personal judgment of the investigator

Data Presentation

Raw data:- It is collected numerical data which has not been arranged in order of magnitude.

An array is an arranged numerical data in order of magnitude.

Data presentation methods:

- Tabular method
- Graphical method
- Diagrammatic method

41

Cont'd

1. Tabular presentation of data:

- The collected raw data should be put into an ordered array in either ascending or descending order so that it can be organized in to a Frequency Distribution (FD)

- Numerical data arranged in order of magnitude along with the corresponding frequency is called frequency distribution (FD).

- FD is of **two kinds: namely ungrouped /and grouped frequency distribution.**

42

Cont'd

A. Ungrouped (Discrete) Frequency Distribution

- ✓ It is a tabular arrangement of numerical data in order of magnitude showing the **distinct values** with the corresponding frequencies.

Example:

Suppose the following are test score of 16 students in a class, write un grouped frequency distribution.

"14, 17, 10, 19, 14, 10, 14, 8, 10, 17, 19, 8, 10, 14, 17, 14"

Sol: the ungrouped frequency distribution:

Array: 8,8,10,10,10,10,14,14,14,14,14, 17,17,17,19,19.

Then the ungrouped frequency distribution is then grouped:-

Test score	8	10	14	17	19
Frequency	2	4	5	3	2

- The difference between the highest and the lowest value in a given set of observation is called **the range** **(R)** **R= L- S, R= 19-8 = 11**

B. Grouped (continuous) Frequency Distribution (GFD)

- ✓ It is a tabular arrangement of data in order of magnitude by classes together with the corresponding class frequencies.

- ✓ In order to estimate the number of classes, the ff formula is used:

Number of classes = 1+3.322(log N) where N is the Number of observation.

$$\text{The Class size (class width)} = \frac{\text{Range}}{1+3.322(\log N)}$$

(round up)

Example:

Grouped/Continuous frequency distribution where several numbers are grouped into one class.

e.g.

Student age	Frequency
18-25	5
26-32	15
33-39	10

46

Components of grouped frequency distribution

1. **Lower class limit:**

   is the smallest number that can actually belong to the respective classes.

2. **Upper class limit:**

   is the largest number that can actually belong to the respective classes.

3. **Class boundaries:**

   are numbers used to separate adjoining classes which should not coincide with the actual observations.

4. **Class mark:**

   is the midpoint of the class.

Cont'd

5. **Class width/ Class intervals**

is the difference between two consecutive lower class limits or the two consecutive upper class limits. (OR)

can be obtained by taking the difference of two adjoining class marks or two adjoining lower class boundaries.

Class width = Range/Number of class desired.

Where: Number of classes=1+3.322(log N) where N is the Number of observation.

6. **Unit of measure**

is the smallest possible positive difference between any two measurements in the given data set that shows the degree of precision.

Cont'd

- ✓ **Class boundaries:**

  can be obtained by taking the averages of the upper class limit of one class and the lower class limit of the next class.

- ✓ **Lower class boundaries:**

  can be obtained by subtracting half a unit of measure from the lower class limits.

- ✓ **Upper class boundaries:**

  can be obtained by adding half the unit of measure to the upper class limits.

Example1 :

Suppose the table below is the frequency distribution of test score of 50 students.

Then the frequency table has 6 classes (class intervals).

Test score	Frequency
11-15	7
16-20	8
21-25	10
26-30	12
31-35	9
36-40	4

What is the Unit of Measure, LCLs, UCLs, LCBs, UCBs, CW, and CM

Cont'd

- ✓ The unit of measure is 1
- ✓ The lower class limits are:-11, 16, 21, 26, 31, 36
- ✓ The upper class limits are:-15, 20, 25, 30, 35, 40
- ✓ The class marks are:- 13((11+15)/2), 18, 23, 28, 33, 38
- ✓ The lower class boundaries are:- 10.5(11-0.5), 15.5, …, 35.5
- ✓ The upper class boundaries are:- 15.5(15+0.5), ….35.5, 40.5
- ✓ Class width (size) is 5.

LECTURE THREE

Rules to construct Grouped Frequency Distribution (GFD):

i. Find the **unit of measure** of the given data;

ii. Find the **range**;

iii. Determine the **number of classes** required;

iv. Find **class width (size)**;

v. Determine a **lowest class limit** and then find the **successive lower and upper class limits** forming non over lapping intervals such that each observation falls into exactly one of the class intervals;

vi. Find the **number of observations** falling into each class intervals that is taken as **the frequency of the class (class interval)** which is best done using a tally.

Exercise:-

Construct a GFD of the following aptitude test scores of 40 applicants for accountancy positions in a company with

a. 6 classes b. 8 classes

96	89	58	61	46	59	75	54
41	56	77	49	58	60	63	82
66	64	69	67	62	55	67	70
78	65	52	76	69	86	44	76
57	68	64	52	53	74	68	39

Types of Grouped Frequency Distribution

1. Relative frequency distribution (RFD)
2. Cumulative Frequency Distribution (CFD)
3. Relative Cumulative Frequency Distribution (RCFD)

Cont'd

1. Relative frequency distribution (RFD):

- A table presenting the ratio of the **frequency of each class to the total frequency of all the classes**.

- Relative frequency generally expressed **as a percentage**, used to show the percent of the total number of observation in each class.

For example

Test score	F	RFD	PFD
37.5-47.5	4	4/40=0.1	10%
47.5-57.5	8	8/40=0.2	20%
57.5-67.5	13	13/40=0.325	32.5%
67.5-77.5	10	10/40=0.25	25%
77.5-87.5	3	3/40=0.075	7.5%
87.5-97.5	2	2/40=0.05	5%

Cont'd

2. Cumulative Frequency Distribution (CFD):

- It is applicable when we want to know how many observations lie **below or above a certain value/class boundary.**

- **CFD** is of two types, **LCFD and MCFD:**
  - ✓ **Less than Cumulative Frequency Distribution (LCFD):** shows the collection of cases lying **below the upper class boundaries of each class.**
  - ✓ **More than Cumulative Frequency Distribution (MCFD):** shows the collection of cases lying **above the lower class boundaries of each class.**

Cont'd

LRCF

Test score	CF
Less than 37.5	0
Less than47.5	4
Less than 57.5	12
Less than 67.5	25
Less than 77.5	35
Less than 87.5	38
Less than 97.5	40

MRCF

Test score	CF
more than 37.5	40
more than 47.5	36
more than 57.5	28
more than 67.5	15
more than 77.5	5
more than 87.5	2
more than 97.5	0

3. Relative Cumulative Frequency Distribution (RCFD)

- It is used to determine the ratio or the percentage of observations that lie below or above a certain value/class boundary, to the total frequency of all the classes. These are of two types: The LRCFD and MRCFD.

- **Less than Relative Cumulative Frequency Distribution (LRCFD):** A table presenting the ratio of the cumulative frequency **less than upper class boundary of each class to the total frequency of all the classes**

- **More than Relative Cumulative Frequency Distribution (MRCFD):** A table presenting the ratio of the cumulative frequency **more than lower class boundary of each class to the total frequency of all the classes.**

Cont’d

LRCFD

Test score	LCF	LRCF	LPCF
Less than 37.5	0	0/40=0	0%
Less than 47.5	4	4/40=0.1	10%
Less than 57.5	12	12/40=0.3	30%
Less than 67.5	25	25/40=0.625	62.5%
Less than 77.5	35	35/40=0.875	87.5%
Less than 87.5	38	38/40=0.95	95%
Less than 97.5	40	40/40=1	100%

Cont'd

MRCFD

Test score	MCF	MRCF	MPCF
More than 37.5	40	40/40=1	100%
More than 47.5	36	36/40=0.9	90%
More than 57.5	28	28/40=0.7	70%
More than 67.5	15	15/40=0.375	37.5%
More than 77.5	5	5/40=0.125	12.5%
More than 87.5	2	2/40=0.05	5%
More than 97.5	0	0/40=0	0%

Graphic Methods of Data presentation

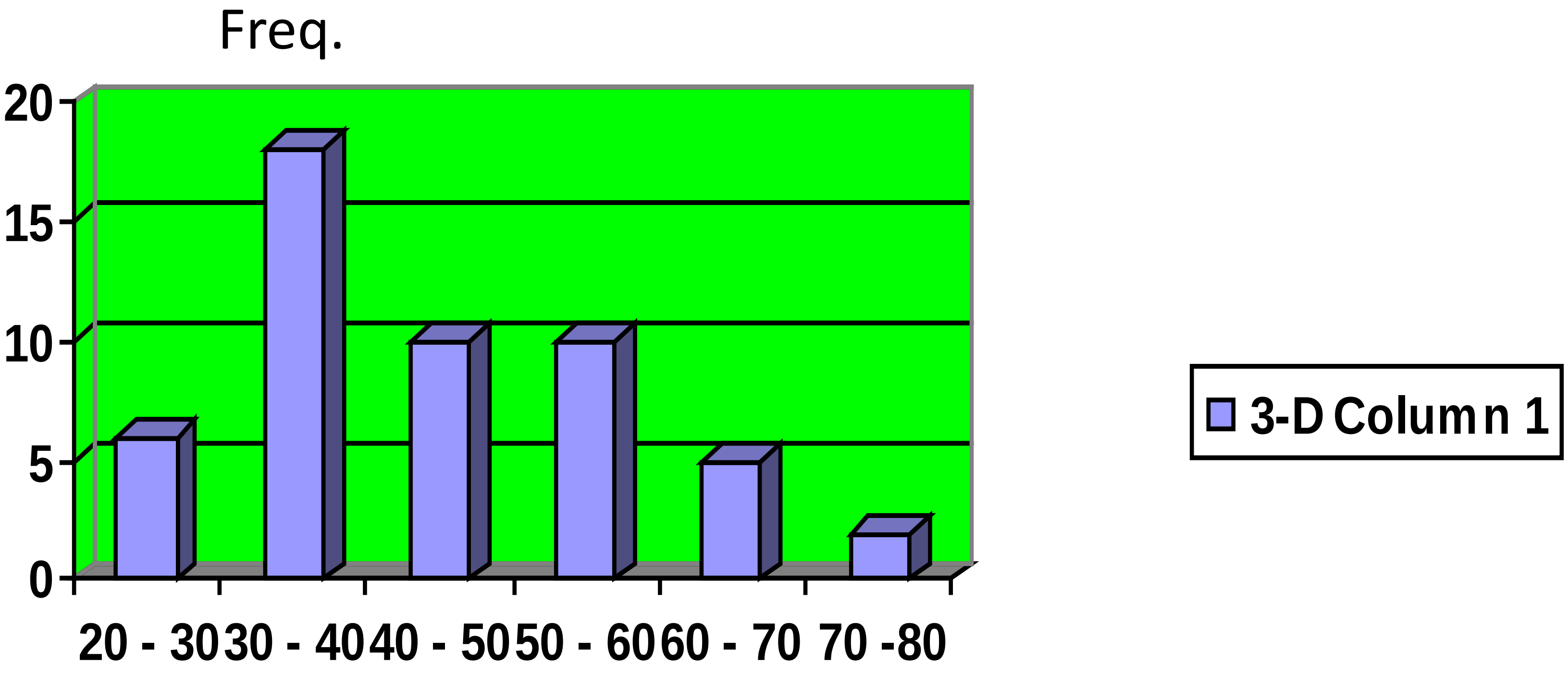
1. Histogram
2. Frequency Polygon (Line graph)
3. Cumulative frequency curve (o-give)

63

1. Histogram:

- A graphical presentation of grouped frequency distribution consisting of a series of adjacent rectangles whose bases are the class intervals specified in terms of class boundaries (equal to the class width of the corresponding classes) shown on the x-axis and whose heights are proportional to the corresponding class frequencies shown on the y-axis.

Histogram: E.g.



Cont’d

Uses for a Histogram

A Histogram can be used:

- to display large amounts of data values in a relatively simple chart form.
- to tell relative frequency of occurrence.
- to easily see the distribution of the data.
- to see if there is variation in the data.
- to make future predictions based on the data.

Steps to draw Histogram

i. Mark the class boundaries on the horizontal axis (x- axis) and the class frequencies along the vertical axis (y- axis) according to a suitable scale.

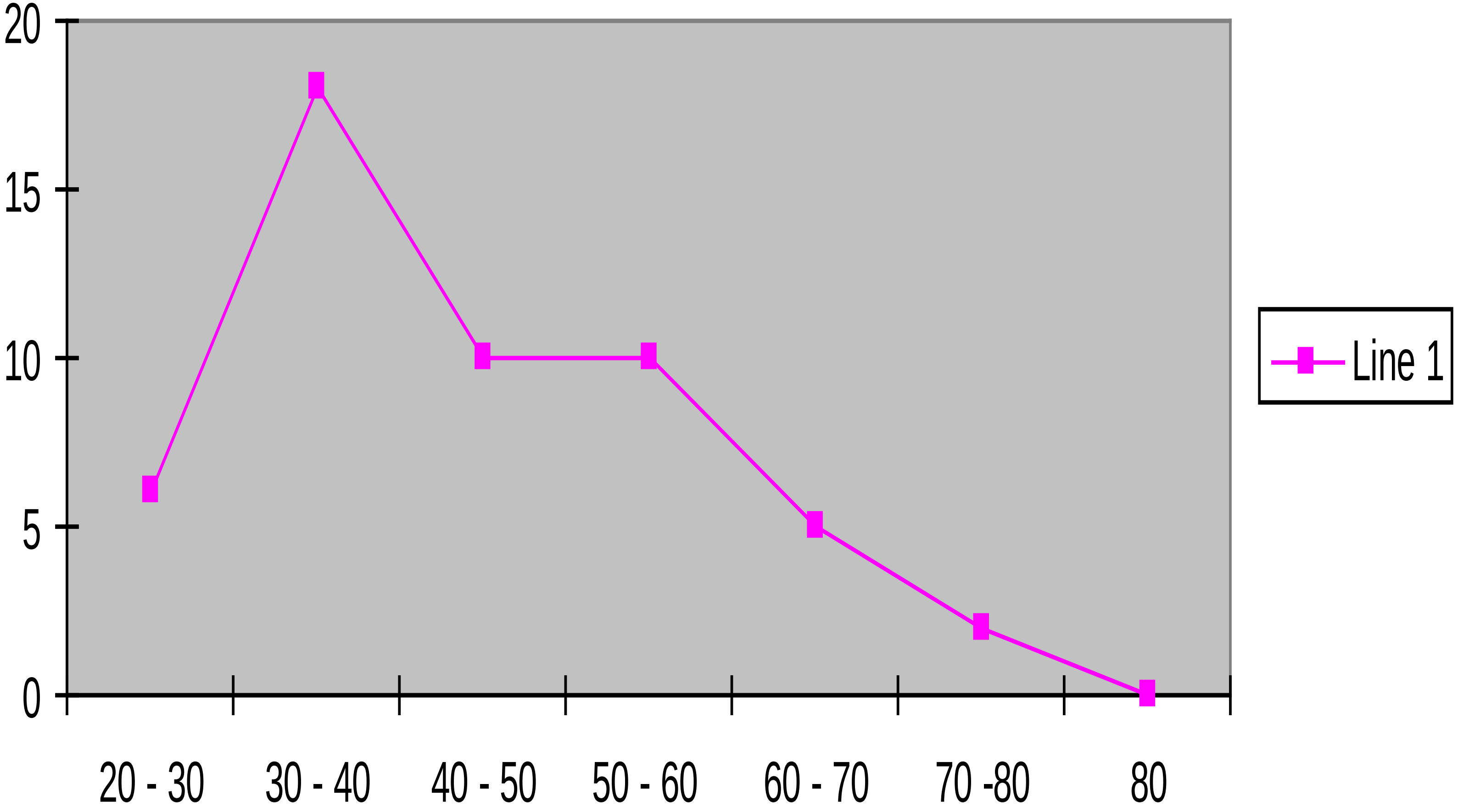
ii. With each interval as a base draw a rectangle whose height equals the frequency of the corresponding class interval. It describes the shape of the data.

2. Frequency Polygon:

It is a line graph of grouped frequency distribution in which the class frequency is plotted against class mark (Midpoint) that are subsequently connected by a series of line segments to form line graph including classes with zero frequencies at both ends of the distribution to form a polygon.

Cont’d

Frequency Polygon:



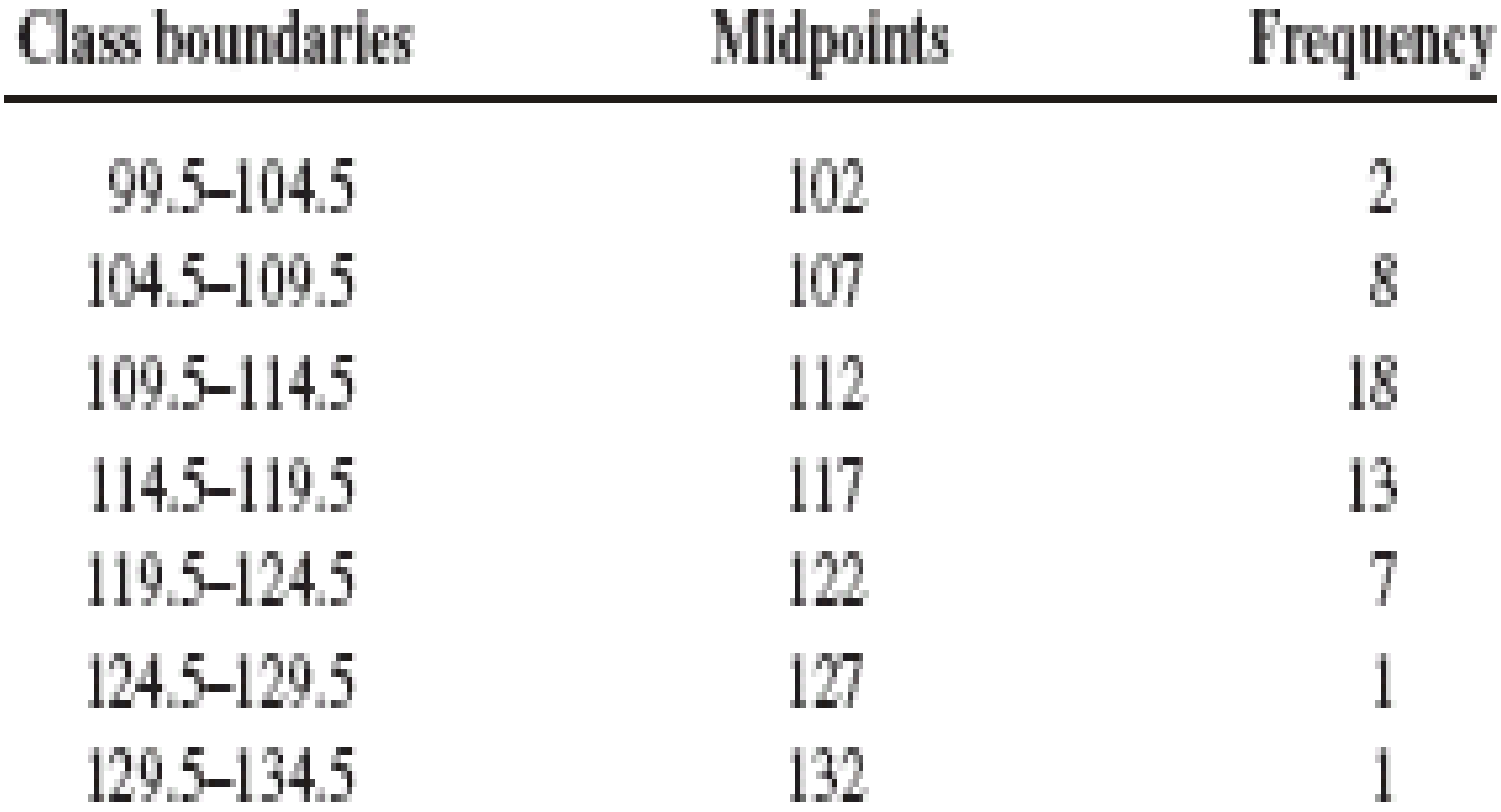
Steps to draw Frequency polygon

i. Mark the class mid points on the x-axis and the frequency on the y-axis.

ii. Mark dots which correspond to the frequency of the marked class mid points.

iii. Join each successive dot by a series of line segments to form line graph, including classes with zero frequencies at both ends of the distribution to form a polygon.

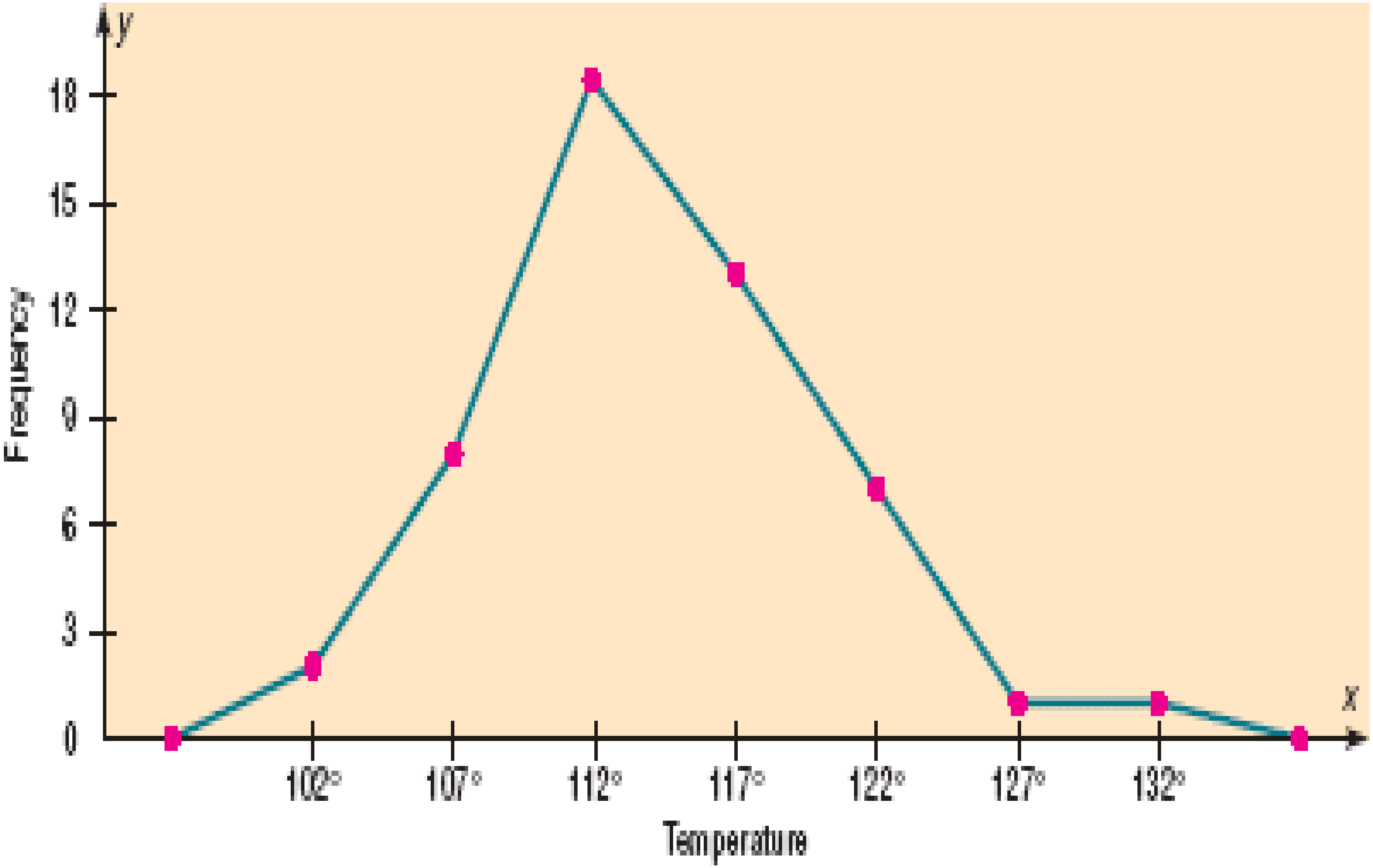
Find the midpoints of each class



Class boundaries	Midpoints	Frequency
99.5–104.5	102	2
104.5–109.5	107	8
109.5–114.5	112	18
114.5–119.5	117	13
119.5–124.5	122	7
124.5–129.5	127	1
129.5–134.5	132	1

Create a frequency polygon using the data

Create a frequency polygon using the data



3. O-GIVE curve (Cumulative Frequency Curve / percentage Cumulative Frequency Curve)

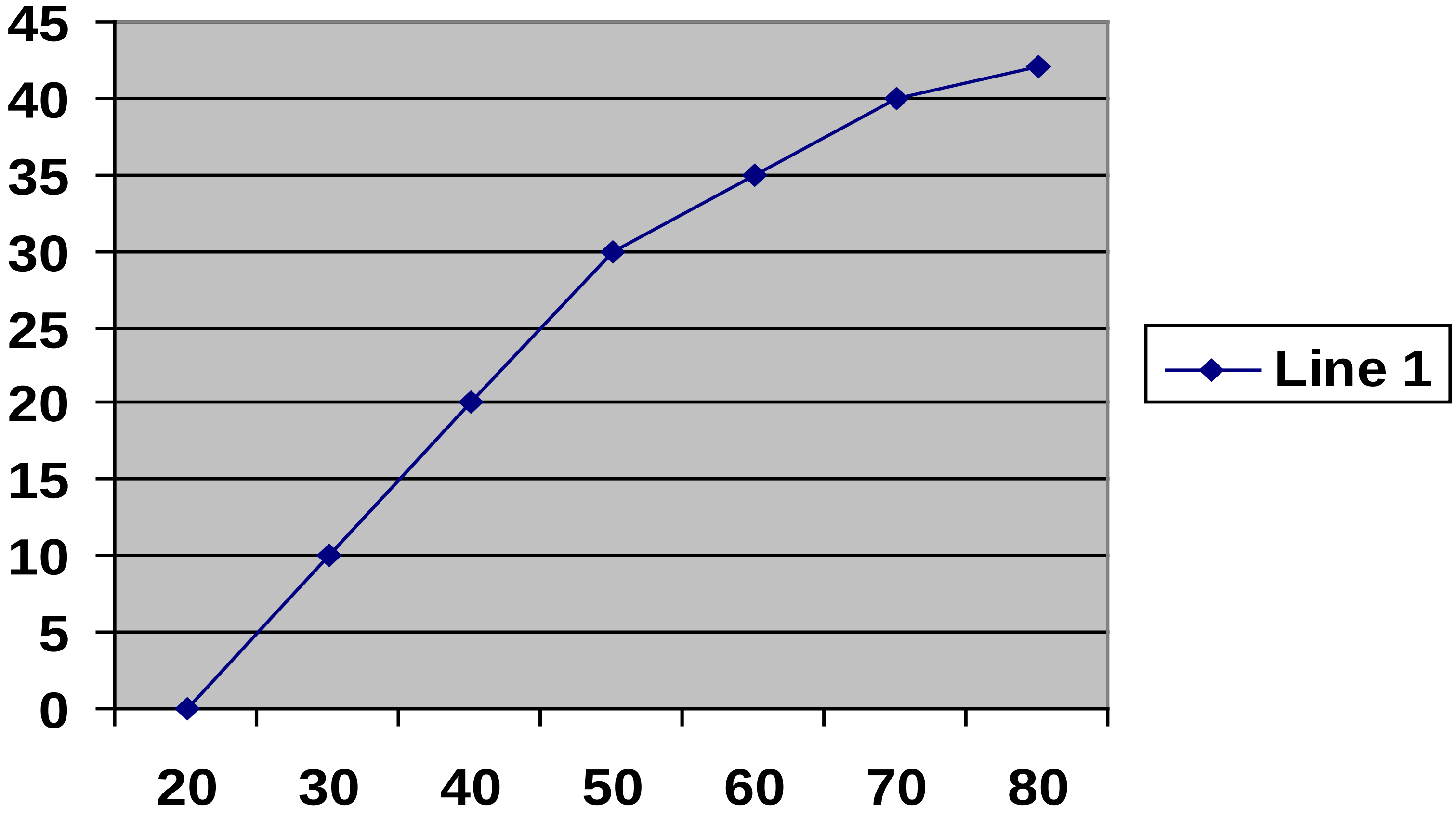
{The number of values less than the upper class boundary for the current class. This is a running total of the frequencies.}

- ✓ It is a line graph presenting the cumulative frequency distribution.
- ✓ O-gives are of two types: The **Less than O-give** and The **More than O-give**.

- ➢ **The Less than O-give** shows the cumulative frequency less than the upper class boundary of each class; and
- ➢ **The More than O-give** shows the cumulative frequency more than the lower class boundary of each class.

Cont'd

Ogive: E.g.



Steps to draw O-gives

i. Mark class boundaries on the x-axis and mark non overlapping intervals of equal length on the y-axis to represent the cumulative frequencies.

ii. For each class boundaries marked on the x-axis, plot a point with height equal to the corresponding cumulative frequencies.

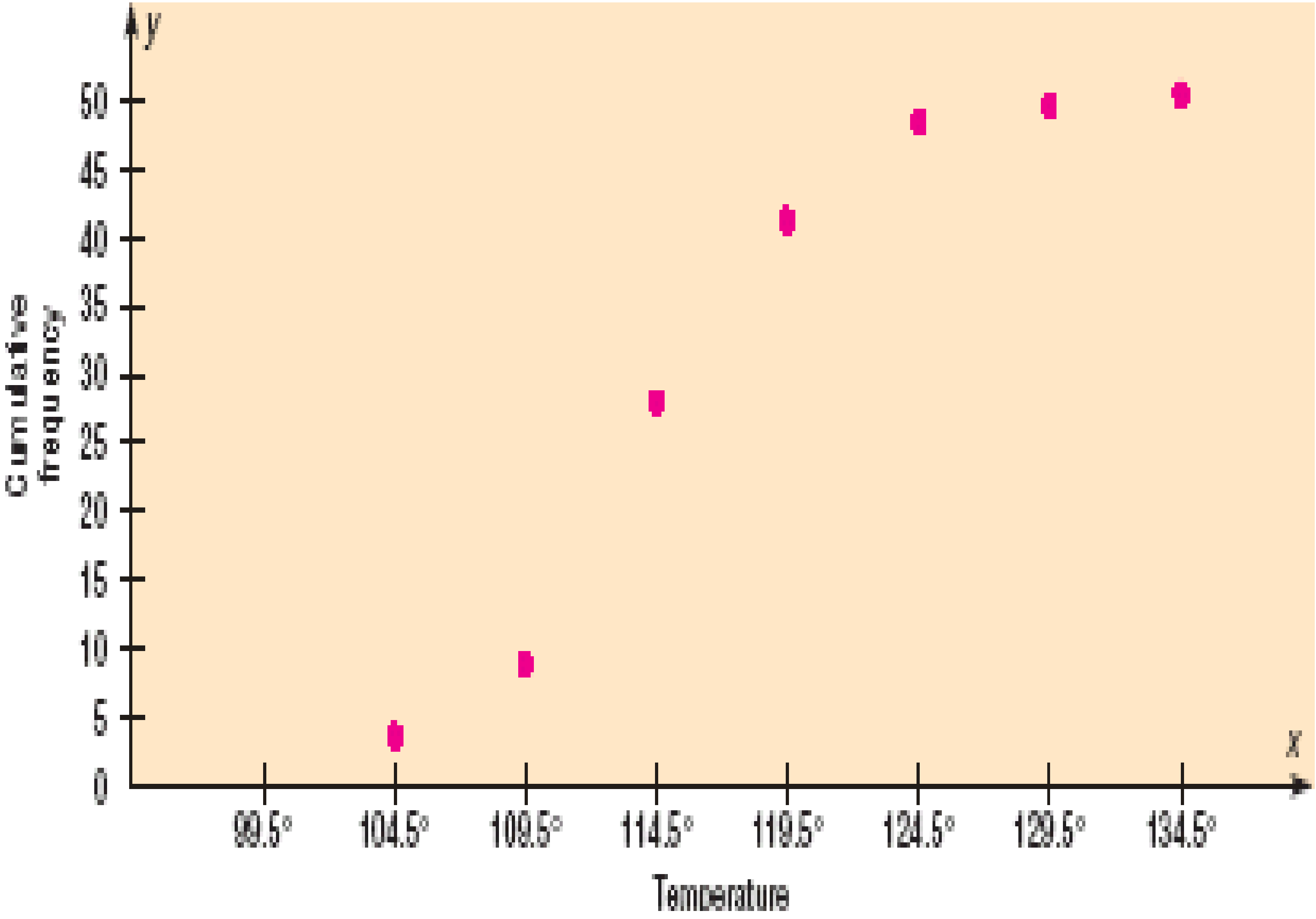
iii. Connect the marked points by a series of line segments where the less than O-give is done by plotting the less than cumulative frequency against the upper class boundaries

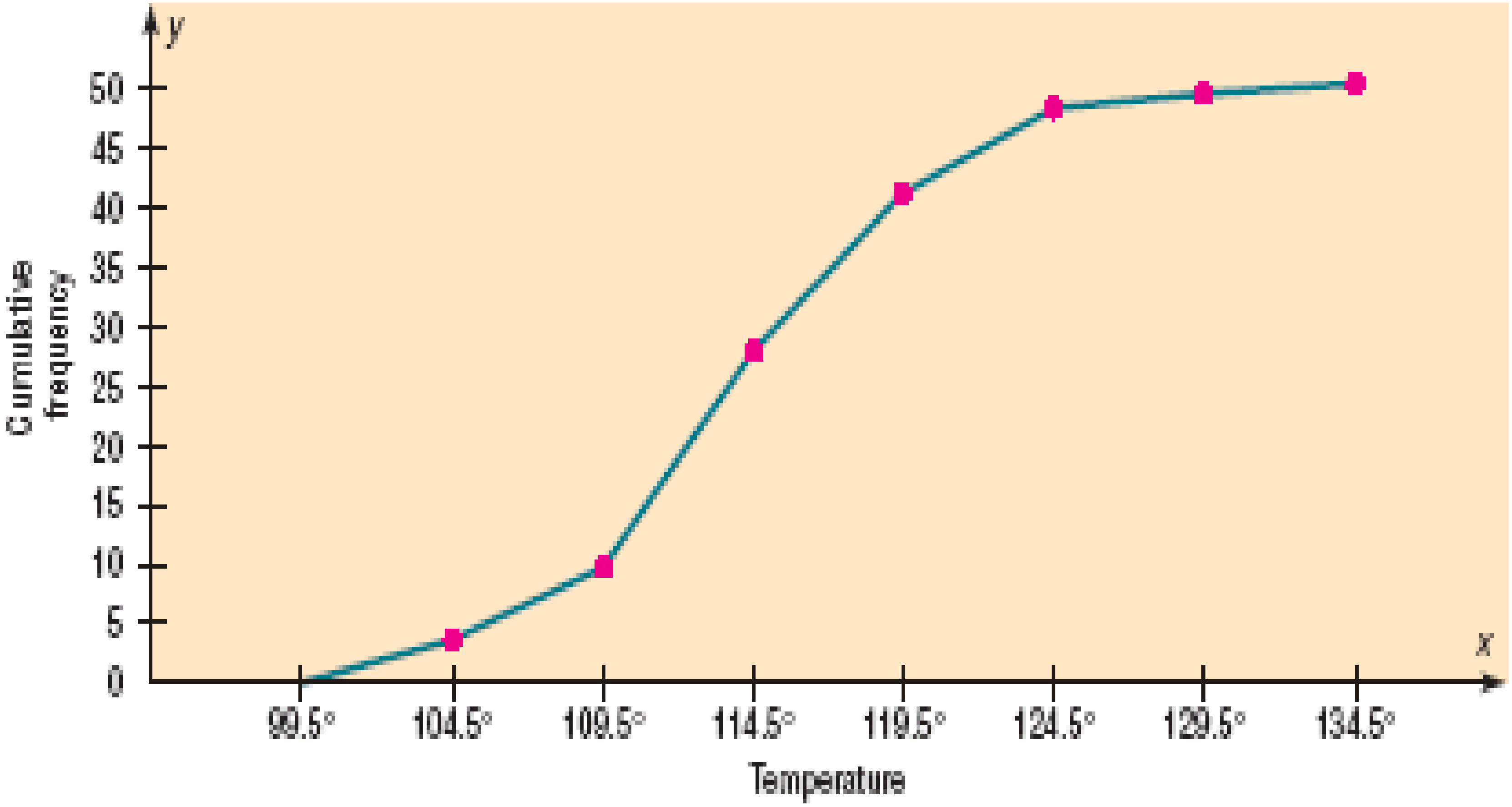
STEP 1 Find the cumulative frequency for each class.

Class boundaries	Cumulative frequency
99.5–104.5	2
104.5–109.5	10
109.5–114.5	28
114.5–119.5	41
119.5–124.5	48
124.5–129.5	49
129.5–134.5	50

Draw the x and y axis

Plot the points


y
50
45
40
35
30
25
20
15
10
5
0
Cumulative frequency
x
99.5°
104.5°
109.5°
114.5°
119.5°
124.5°
129.5°
134.5°
Temperature


y
50
45
40
35
30
25
20
15
10
5
0
Cumulative frequency
x
99.5°
104.5°
109.5°
114.5°
119.5°
124.5°
129.5°
134.5°
Temperature

Diagrammatic Presentation Of Data

- Bar charts
- Pie chart
- Pictograph and
- Pareto diagram

Outline

1.3 Measures of Central Tendency

1.4 Measures of Dispersion

1.5 Measure of skewness and kurtosis

1.3 Measures of Central Tendency

- ✓ Central value refers to the location of the centre of the distribution of data.
- ✓ Measure of central value:
  - ➢ **The mean**
  - ➢ **The mode**
  - ➢ **The median**
  - ➢ **Percentile/Quantiles and**
  - ➢ **Midrange**

The Mean:

- ✓ It is the most commonly used measures of central value.
- ✓Types of Mean:
  1. Arithmetic Mean
  2. Geometric Mean
  3. Harmonic Mean
  4. Quadratic Mean
  5. Trimmed Mean
  6. Weighted Mean
  7. Combination mean

1. Arithmetic Mean (simply **Mean**)

✓ The mean is defined as the arithmetic average of all the values.

✓ It is represented by $\bar{x}$ **read as x-bar for a sample and by μ for a population**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Cont'd

Advantages

- It is the most commonly used measure of location or central tendency for continuous variables.
- The arithmetic mean uses all observations in the data set.
- All observations are given equal weight.

Disadvantage

- The mean is affected by **extreme values** that may not be representative of the sample.

2. Geometric mean

- It is the ***n*th root** of the product of the data elements.
- It is used in business to find average rates of growth.

Geometric mean = $\sqrt[n]{\prod x_i}$,

for all *n* >=2.

Cont’d

- **Example:** suppose you have an IRA (Individual Retirement Account) which earned annual interest rates of 5%, 10%, and 25%.

  Solution: The proper average would be the geometric mean or the cube root of (1.05 • 1.10 • 1.25) or about 1.13 meaning 13%.

- Note that the data elements must be positive. Negative growth is represented by positive values less than 1. Thus, if one of the accounts lost 5%, the proper multiplier would be 0.95.

3. Harmonic mean

✓It is used to calculate average rates.

✓It is found by dividing the number of data elements by the sum of the reciprocals of each data element.

Harmonic mean = $\frac{n}{\sum x_i^{-1}}$

- **Example:** Suppose a boy rode a bicycle three miles. Due to the topography, for the first mile he rode 2 mph; for the second mile 3 mph; for the final mile the average speed was 4 mph. What was the average speed for the three miles?

Solution:

To show simple analysis using Harmonic mean :

$$\frac{3 \text{ miles}}{1/2+1/3+1/4} = \frac{3 \quad .}{13/12} = \frac{36}{13} = 2.77\text{mph}$$

4. Quadratic mean

- It is another name for **Root Mean Square or RMS.**
- **RMS** is typically used for data whose arithmetic mean is zero.

$$\mathrm{RMS} = \sqrt{\frac{\sum x_i^2}{n}}$$

Cont’d

Example

- Suppose measurements of 120, -150, and 75 volts were obtained.

Solution: The corresponding quadratic mean is √((120)2 + (-150)2 + (75)2)/3) or 119 volts RMS.

- The quadratic mean gives a physical measure of the average distance from zero.

5. Trimmed mean

- It is usually refers to the arithmetic mean without the top 10% and bottom 10% of the ordered scores.

- Removes extreme scores on both the high and low ends of the data

6. Weighted mean

✓ It is the average of differently weighted scores;

✓ It takes into account some measure of weight attached to different scores.

$$\text{Weighted mean} = \frac{\sum (w_i \cdot x_i)}{\sum w_i}$$

The Mode

- The mode is the most frequent or most typical value.
- The mode will not always be the central value; in fact it may sometimes be an extreme value. Also, a sample may have more than one mode **Bimodal** or ***Multimodal***

Cont'd

Example

'23, 22, 12, 14, 22, 18, 20, 22, 18, 18'

The mode is 18 and 22 - bimodal

Cont'd

Advantages

- Requires no calculations.
- Represents the value that occurs most often.

Disadvantage

- The mode for continuous measurements is dependent on the grouping of the intervals.
- We may not have mode

The Median

- The median is the middle value of a group of an odd number of observations when the data is arranged in increasing or decreasing order magnitude.

- If the number of values is even, the median is the average of the two middle values.

Cont'd

Example

'23, 22, 12, 14, 22, 18, 20, 22, 18, 18'

Array: 12, 14, 18, 18, 18, 20, 22, 22,22,23

- The median (18+20)/2 = 19

Cont'd

Advantages

- The median always exists and is unique.
- The median is not affected by extreme values.

Disadvantages

- The values must be sorted in order of magnitude.
- The median uses only one (or two) observations.

Percentiles / Quartiles

- Percentiles are values that divide a distribution into two groups where the **Pth** percentile is larger than **P%** of the values.

- Some specific percentiles have special names:

First Quartile : Q_1 = the 25 percentile

Median : Q_2 = the 50 percentile

- A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.

Cont'd

Interpretation

- Q_1 = a. This means that 25% of the data values are smaller than a
- Q_2 = b. This means that 50 % of the data has values smaller than b.
- Q_3 = c. This means that 75 % of the data has values smaller than c.

Midrange

- The midrange is the average of largest and smallest observation.

Midrange = (Largest +Smallest)/2

- The percentile estimate (P25 + P75)/2 is sometimes used when there are a large number of observations.

LECTURE FOUR

1.4 Measure of Dispersion (of Variability)

- So far we have looked at ways of summarising data by showing some sort of average (central tendency).

- But it is often useful to show how much these figures differ from the average.

- This measure is called **dispersion.**

Cont'd

- *Measures of dispersion* are descriptive statistics that describe how similar a set of scores are to each other.
  - ❖The more similar the scores are to each other, the lower the measure of dispersion will be.
  - ❖The less similar the scores are to each other, the higher the measure of dispersion will be
  - ❖In general, the more spread out a distribution is, the larger the measure of dispersion will be.

Cont'd

- A measure of dispersion indicates how the observations are spread about the central value.

Measures of dispersion are:

- **The range**
- **The variance**
- **The standard deviation and**
- **The coefficient of variation**

The range

- The range is the difference between the largest and the smallest value in the sample.
- The range is the easiest of all measures of dispersion to calculate.

R = Maximum Value - Minimum Value.

Cont'd

Advantage

- The range is easily understood and gives a quick estimate of dispersion.
- The range is easy to calculate

Disadvantage

- The range is **inefficient** because it only uses the extreme value and ignores all other available data. The larger the sample size, the more inefficient the range becomes.

The Variance

The variance is the average of the squared differences between each data value and the mean.

If the data set is a sample, the variance is denoted by s^2.

$$\boxed{s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}}$$

If the data set is a population, the variance is denoted by σ^2

$$\boxed{\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}}$$

Cont'd

Advantages

- The variance is an efficient estimator
- Variances can be added and averaged

Disadvantage

- The calculation of the variance can be tedious without the aid of a calculator or computer.

The Standard Deviation

- The standard deviation is one of the most important measures of dispersion. It is much more accurate than the range or inter quartile range.
- It takes into account all values and is not excessively affected by extreme values.
- The square root of the variance is known as the standard deviation.
- If the data set is a sample, the standard deviation is denoted *s*.

$$\boxed{s = \sqrt{s^2}}$$

- If the data set is a population, the standard deviation is denoted by σ

$$\boxed{\sigma = \sqrt{\sigma^2}}$$

Cont'd

What does it measure?

- It measures the dispersion (or spread) of figures around the mean.

- A large number for the standard deviation means there is a wide spread of values around the mean, whereas a small number for the standard deviation implies that the values are grouped close together around the mean.

Example

- Find the standard deviation of the minimum temperatures of 10 weather stations in Ethiopia on a winters day.

The temperatures are:

5, 9, 3, 2, 7, 9, 8, 2, 2, 3 (°Centigrade)

Solution

x	$\ddot{x}$	$(x - \ddot{x})$	$(x - \ddot{x})^2$
5	5	0	0
9	5	4	16
3	5	-2	4
2	5	-3	9
7	5	2	4
9	5	4	16
8	5	3	9
2	5	-3	9
2	5	-3	9
3	5	-2	4
$\sum x =$ <u>50</u> $\ddot{x} = \sum x/n =$ <u>50/10 = 5</u>			$\sum(x - \ddot{x})^2 =$ <u>80</u> $\sum(x - \ddot{x})^2/N =$ <u>8</u> $\sqrt{\sum(x - \ddot{x})^2/N} =$ 2.8°C

x = temperature --- $\ddot{x}$ = mean temperature --- $\surd$ = square root

$\sum$ = total of --- ^{2} = squared --- n = number of values

The coefficient of variation

- The coefficient of variation indicates how large the standard deviation is in relation to the mean.
- If the data set is a sample, the coefficient of variation is computed as follows:

$$\boxed{\frac{s}{\overline{x}}(100)}$$

- If the data set is a population, the coefficient of variation is computed as follows:

$$\boxed{\frac{\sigma}{\mu}(100)}$$

Cont'd

- The **higher** the Coefficient of Variation the **more widely** spread the values are around the mean.

- The purpose of the Coefficient of Variation is to let us compare the spread of values between different data sets.

The Inter-Quartile Range

- It is the difference between the 25th and the 75th quartiles.
- The inter-quartile range is the range of the middle half of the values.
- It is a better measurement to use than the range because it only refers to the middle half of the results.
- Basically, the extremes are omitted and cannot affect the answer.

Interquartile range = Q3 – Q1

Cont'd

- To calculate the inter-quartile range we must first find the **quartiles**.

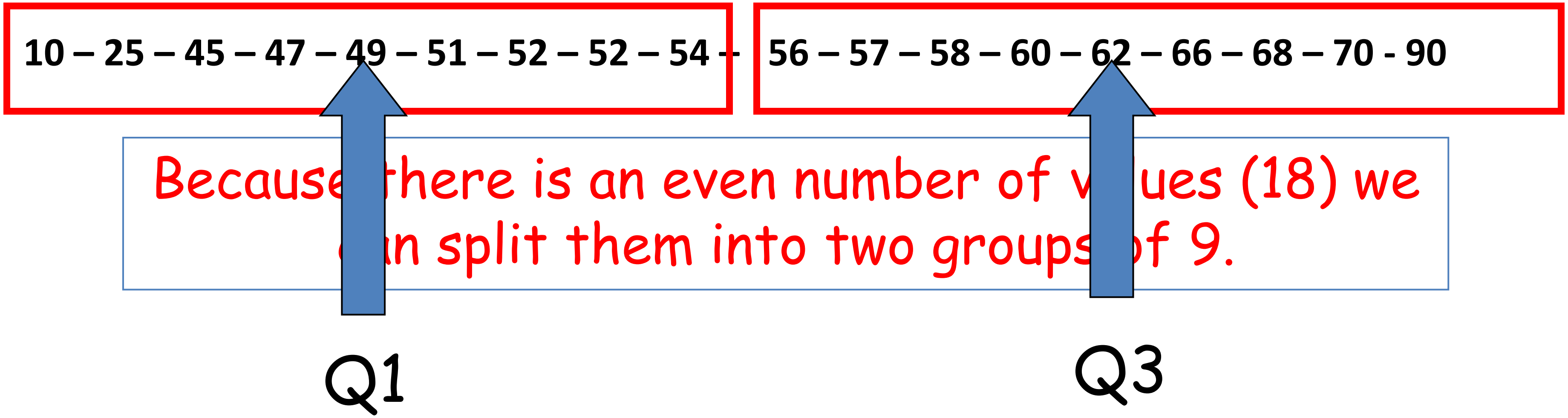
- There are three quartiles, called Q1, Q2 & Q3. We do not need to worry about Q2 (this is just the median).

- Q1 is simply the middle value of the bottom half of the data and Q3 is the middle value of the top half of the data.

Example

- We calculate the **inter quartile** range by taking Q1 away from Q3 (Q3 – Q1).

Remember data must be placed in order



10 – 25 – 45 – 47 – 49 – 51 – 52 – 52 – 54 –

56 – 57 – 58 – 60 – 62 – 66 – 68 – 70 - 90

Because there is an even number of values (18) we can split them into two groups of 9.

IR = Q3 – Q1 , IR = 62 – 49. IR = 13

Measure of skewness and kurtosis

Skewness (mean deviation)

- Skewness is a measure of the tendency of the deviations to be larger in one direction than in the other.
- Skewness is the degree of asymmetry or departure from symmetry of a distribution.
- If the frequency curve of a distribution has a longer tail to the right of the central value than to the left, the distribution is said to be **skewed to the right or to have positive skewness.**
- If the reverse is true, it is said that the distribution is **skewed to the left or has negative skewness**.

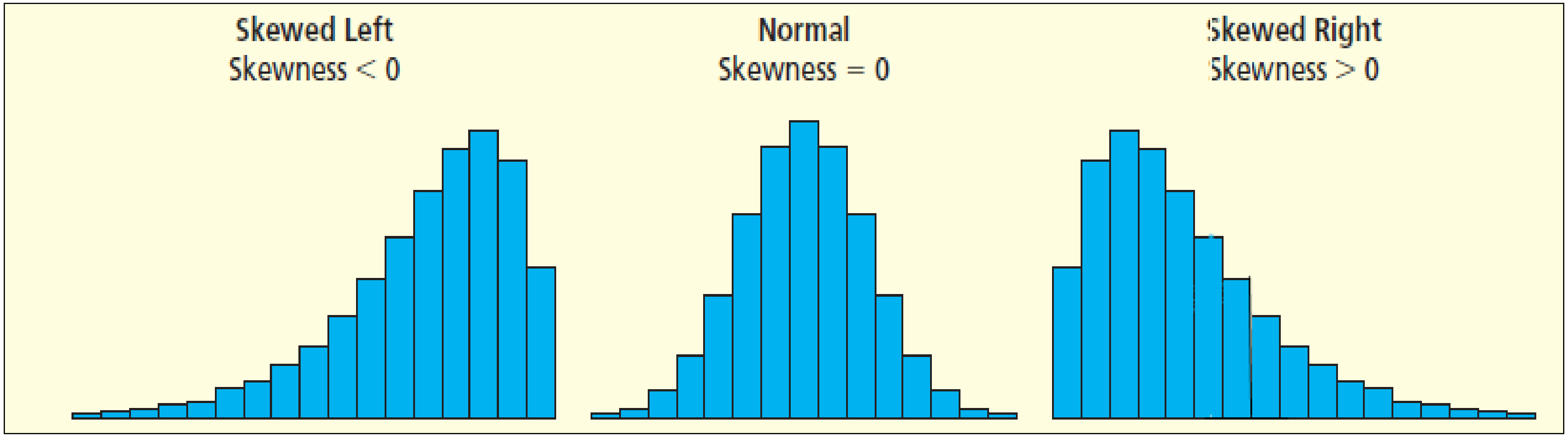
Cont'd

The following formula can be used to determine skew:

$$\boxed{S^3 = \frac{\dfrac{\sum\left(X-\overline{X}\right)^3}{N}}{\sqrt{\dfrac{\sum\left(X-\overline{X}\right)^2}{N}}}}$$

- A bell-shaped distribution which has no skewness, i.e., **mean = median = mode** is called a normal distribution. If **mean > Median > mode**, the distribution is positively skewed distribution or it is said to be skewed to the right. If mean < Median < mode, the distribution is negatively skewed distribution or it is said to be skewed to the left.

Cont'd



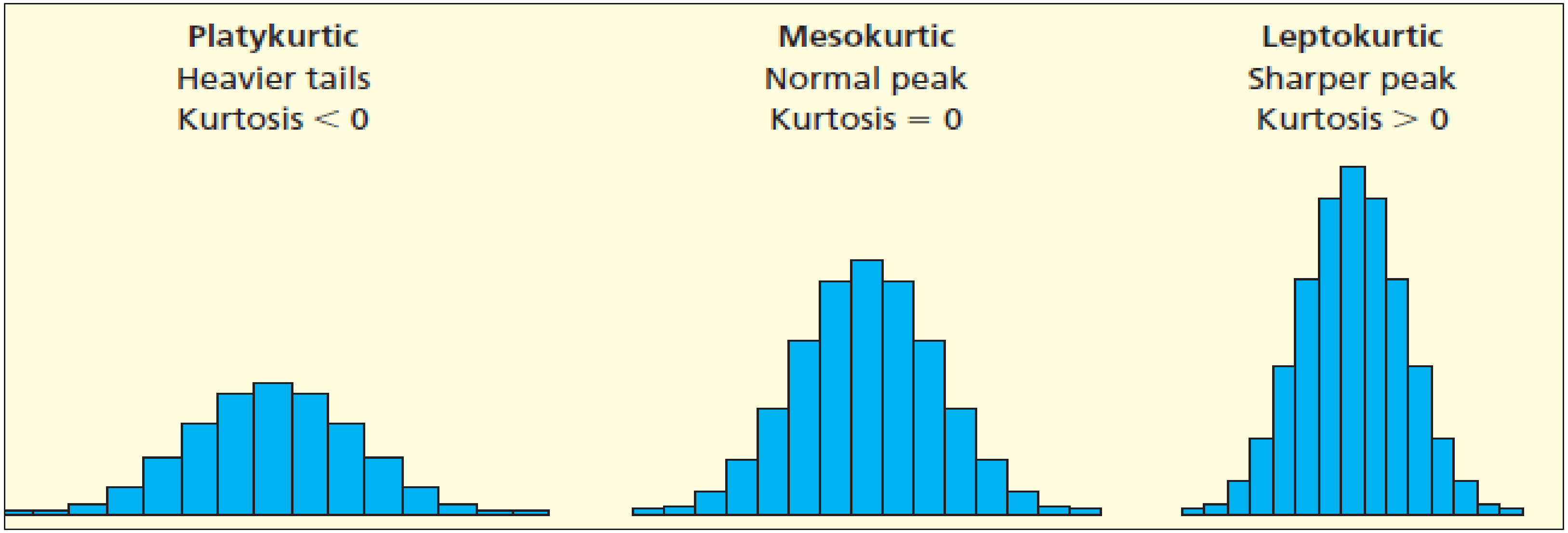
- If $s^3 < 0$, then the distribution has a negative skew
- If $s^3 > 0$ then the distribution has a positive skew
- If $s^3 = 0$ then the distribution is symmetrical
- The more different s^3 is from 0, the greater the skew in the distribution

Measure of kurtosis

- Kurtosis characterises the relative peakedness or flatness of a distribution compared with the normal distribution. Positive kurtosis indicates a relatively peaked distribution. Negative kurtosis indicates a relatively flat distribution.

The measure of kurtosis is given by:

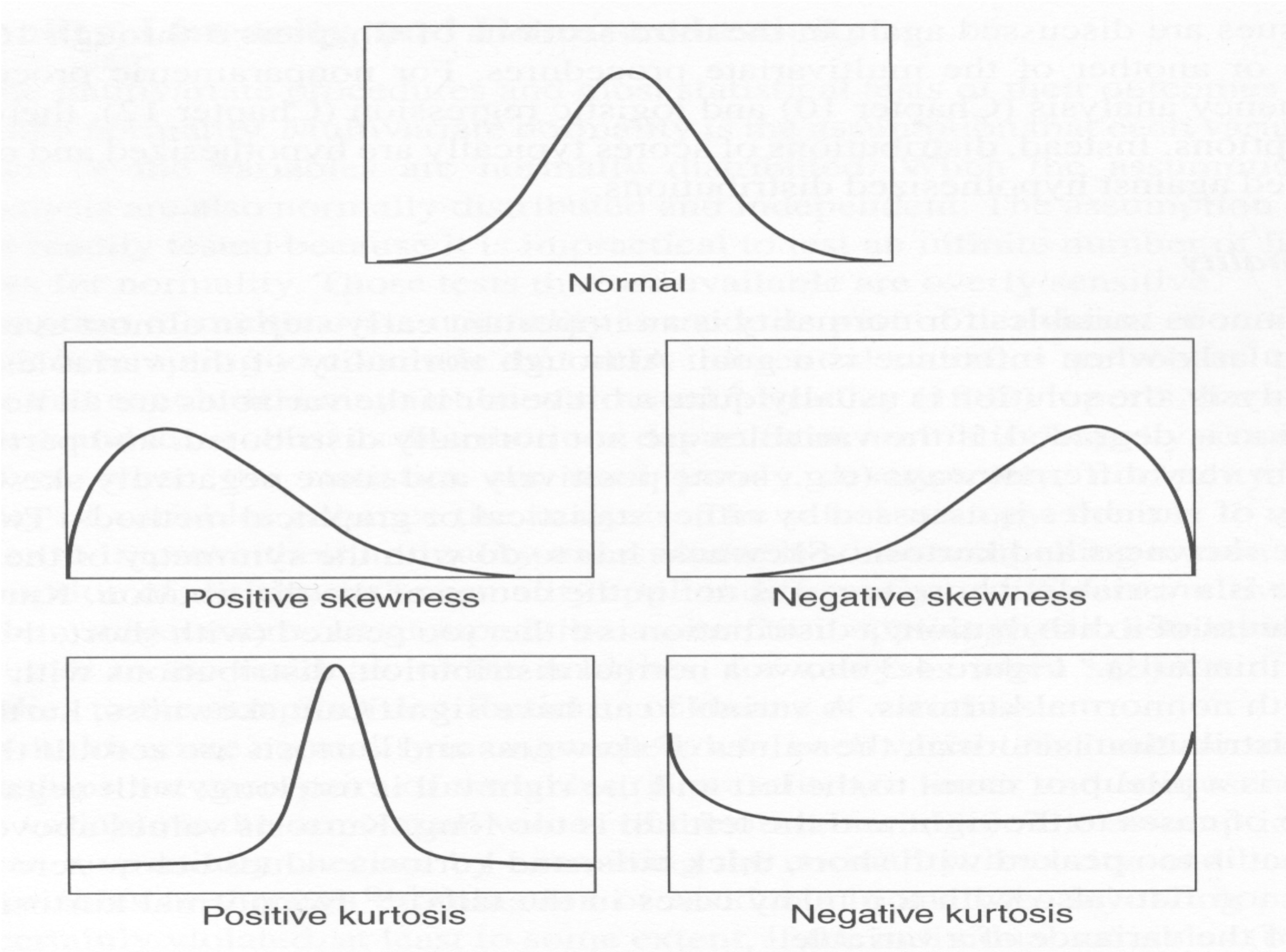
$$s^4 = \frac{\sum \left(\frac{X - \overline{X}}{\sqrt{\frac{\sum \left(X - \overline{X} \right)^2}{N}}} \right)^4}{N}$$



s^2, s^3, & s^4

Collectively, the variance (s^2), skew (s^3), and kurtosis (s^4) describe the shape of the distribution

Skewness and Kurtosis



End of Chapter 1